Modelling and validation techniques for bottom-up housing stock modelling of non-heating end-use energy in England

Stephen William Lorimer

A thesis submitted for the degree of Doctor of Philosophy

University College London

2012

Declaration

I, Stephen William Lorimer, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

_____

Abstract

This thesis engages with different methods and validation techniques for bottom-up stock modelling of non-heating end-use energy of the residential sector. These end-uses are not the primary focus of current domestic energy models, and there is a unique opportunity to use actual electricity use data to build and validate models as electricity becomes exclusively used for these end-uses in England.

The first contribution to knowledge is the creation of a validation set from aggregated electricity use data that has become available from small census areas of around 600 households using only areas with minimal estimated rates of electric heating. The second contribution is a method for using partial data from recent housing and energy surveys to update complete, but dated surveys by using household size and seasonal distributions. This enables a yearly updated model validated against actual aggregate energy use.

This led to an annually updateable single-level model of non-heating end-use energy based on the predictors of household size measured by the number of rooms and the number of occupants. This uses linear regression on a square-root transformation of energy instead of the current natural logarithm transformation. The model is found to have a slight over-prediction (1.5%) of energy use when validated.

The final contribution is an alternative approach where the model was allowed to vary on the household's area. A hierarchical linear model of domestic energy was built based on 20 area classifications. There is a weak, but significant effect of additional energy use in households located inside area classifications with higher mean household sizes. This effect is highly significant when building age is taken into account. Although validation was difficult because building age data is limited, this result points to a neighbourhood-level influence that explains energy use beyond individual household size if precise location data can be made available.

Acknowledgements

I would like to first pay tribute to my supervisor, Harry Bruhns, who sadly passed away before he could see the fruits of his labour in this thesis. He will be missed by all who have come in contact with him.

Many thanks go to my supervisors Bob Lowe, building physicist, and without his questioning and support for work out of my comfort zone I could not have completed this thesis, and to Phil Steadman, architect and polymath, who always was there to give me confidence in my ability.

There was significant financial support provided to me from the UCL Energy Institute via the Engineering and Physical Research Council Doctoral Training Account grant, without which this work would not have been possible.

A word of thanks should also go Terry McIntyre, Principal Researcher at the Department for Communities and Local Government, who permitted me to match location data in the form of area classifications to the homes surveyed in the 1996 English House Condition Survey. It proved to be a unique and invaluable dataset for this research.

Table of Contents

Table of Figures

List of Tables

Chapter 1 - Introduction

1.1 Introduction

This PhD thesis develops methods for estimating non-heating end-use energy for the existing residential sector in all the districts of England. Non-heating end-use energy is defined as the electricity delivered to the residential sector for end-uses not associated with space or water heating. The estimates are developed for the year 2001 and are projected through to 2008 using additional data.

This thesis engages with different methods and validation techniques for bottom-up stock modelling of non-heating end-use energy of the residential sector. These end-uses are not the primary focus of current domestic energy models because they are based on the heat flux between the inside and outside of a building as heat demand takes up by far the largest part of all energy demand. However, over the last 30 years this has gradually changed. Measuring, modelling, and reducing energy demand for heating has been a singular focus of the built environment research community mapping out predictors ranging from materials, technologies, and weather to demand. There has been considerable progress made in this area with a number of factors researched and discovered compared to energy demand for lights, appliances, and electronics, which have roughly doubled their demand in this time period. These end-uses are less dependent on variations in installed technologies, materials science, and weather – instead, according to different models, they are dependent on household size or the market demand and natural wastage of consumer goods.

Decarbonisation of the energy supply in the UK will likely involve the transfer of heat energy demand from gas to electricity. This would, at present, necessitate a massive increase in both generation and transmission capacity. As electricity use continues to increase from other end-uses, there may be incorrect assumptions made in housing stock models about the electricity capacity available for heating in the future as this capacity is already taken up, delaying transfer of heating fuels and therefore delaying decarbonisation of the energy supply.

The demand for information on non-heating end-use energy stems from important information gaps. Bottom-up domestic stock models built up from the predicted energy consumption of homes as designed and then validated against a second model of appliance ownership. In this thesis, a stock

model is built to predict consumption as occupied, and then validated against the recent history of electricity use aggregated by area typology. This validation set is available for small census areas of around 600 households with evidence that there was negligible use of electricity for space and water heating.

The main methodological contributions of this thesis are two methods for estimating the non-heating end-use energy of a residential district. The first is a method for using partial data to project estimates made from 10-yearly censuses to more recent years, therefore updating the stock model of electricity use of dwellings. This model of non-heating end-use energy uses individual household sizes of existing homes as a predictor of energy use of households that were the building blocks of a bottom-up energy model of the residential sector, validated against the aggregate totals for districts. The second method allows these districts to vary and to be a predictor themselves, resulting in a model that predicts non-heating energy use in households with the characteristics of both households and districts simultaneously.

This introductory chapter sets out the research context and is divided into 5 parts. The first part argues that domestic stock modelling of non-heating end-use energy, as distinct from the modelling of domestic heating demand, is important and will become more so with the decarbonisation of energy supplies. The second examines the information gap being pursued in this thesis – that stock models of non-heating end-use energy are not validated against actual electricity use. The third section discusses techniques discovered for annual updates of a model using partial data. The fourth section describes the creation of a model using household size as a predictor that feeds into this method of updating and why it is different from the current model. The fifth section explains the creation of multi-level model predicting energy use using both households and districts. Finally, the structure of the thesis is set out.

1.2 Why domestic stock modelling of non-heating end-use energy is important

Domestic stock modelling of non-heating end-use energy is important in the context of energy use in England because demand continues to grow, there are plans to move heating fuel from natural gas to electricity whilst the present gasification of heating fuel means that non-heating end-uses are easily measured in the present. Firstly, there is a need to understandi and control non-heating end-uses of electricity in its own right as demand for non-heating end-uses has grown at a faster pace in the developed world than for heating. Although there are plans to decarbonise electricity generation, rising demand in non-heating end-uses could slow or, in the worst case scenario, even

block the transfer of heating demand onto the electricity grid. As it has been cheaper to directly burn fossil fuels in the home for heating, electricity is little-used as a heating fuel, making England a good laboratory for the investigation of this type of energy demand. This leads to the conclusion that it is important to investigate domestic energy stock modelling of non-heating end-use energy.

Demand for non-heating end-use energy is still growing and is predicted to continue to do so with most energy conservation and efficiency measures focussed on heating. Recent data from the current Housing Energy Fact Files for the United Kingdom reveals that from 1970 to 2006, the energy consumption of lighting and appliances grew at around seven times the rate of the overall growth in energy use (Shorrock and Utley, 2008, Department of Energy and Climate Change, 2011). The rate of increase will rise from 1.25% per year over the last thirty years to 5% per year over the next twenty years. Taking just electronic gadgets, electricity use is predicted to rise by an average of 12.5% per year over the next forty years (International Energy Agency, 2009, Ekins and Dresner, 2006). There are more modest projections that are currently being put in place for UK government modelling, including the rate of increase to remain the same to 2050 (HM Government, 2010). There is little will to address this demand in the built environment – most of the effort, rightly, is on reductions in the majority of energy demand as heating, and even less in the political realm as objects and gadgets are much more a part of one's feeling of well-being and status within a modern society in a developed country than the window glazing or wall insulation of a building.

The move to "decarbonised electricity" as a heating fuel will place enormous pressure on the electricity grid to deliver the required energy for both heating and non-heating end-uses. Even if heating demand is greatly reduced in the future, the doubling, or in the case of the pessimistic outlook of the IEA, the quintupling, of non-heating demand will require electric generation capacity far greater than imagined using current energy use patterns. In other words, raised non-heating energy demand will retard or even prevent a switch from natural gas to electric fuels for heating should the decarbonisation of the electricity supply come to fruition.

This time is particularly fertile in England, and the rest of the United Kingdom, for the study of usage patterns of appliances, lighting, cooking, and electronics that make up non-heating end-use energy. Electric heating is used in around 10% of dwellings in the UK. Examples can be found in electric storage heating in Scotland, modern apartment blocks unconnected to the gas grid due to modern gas safety regulations, and housing in city centres such as Manchester without a connection to the natural gas network. Britain has developed its energy supply for heating with a natural gas resource that has been historically abundant in Britain's North Sea oil fields since the 1970s.

This situation is likely only to exist in a brief window of time because of UK government policy on zero carbon housing (Department for Communities and Local Government, 2008a). The decarbonisation of the housing stock using existing technologies is likely to involve a transfer of heating fuel from natural gas to electricity generated from clean energy sources. Some estimates for decarbonising energy call for a reduction of the ownership level of natural gas boiler ownership to achieve a 40 percent reduction in total carbon emissions from 1990 levels (Boardman, 2005, Johnston et al., 2005). This drastic change attracted some criticism as it implies that many existing homes were thermally "irredeemable" (Lowe, 2007). Another analysis led by Imperial College states that meeting the UK Government's emissions and renewables targets leads to a higher peak load of up to 92% if electrification of transport and heating goes ahead (Imperial College and Energy Networks Assocation, 2010).

These three reasons, of rising demand, the difficulties this imposes on the decarbonisation of heating demand, and the situation in the UK where these end-uses are easily measurable, lead to the conclusion that this is an important area of study within energy research.

1.3 Information gap - Validation set

This thesis identifies an information gap in the validation of domestic energy stock modelling of non-heating end-use energy. The current validation set is a second model called the DECADE model, developed at the Oxford University Environmental Change Institute during the 1990s (Environmental Change Institute, 1995). This model was designed as a large conditional demand analysis model, with the total number of appliances owned multiplied by their usage hours and consumption rate to estimate total energy use for non-heating end-use energy. This technique was initially developed to estimate the electricity use of individual households, but the DECADE model created a large model of all appliances that were in use in the United Kingdom, the appliance or electronics type, their vintage (year bought), and the rate of energy use of that appliance of a certain vintage.

The domestic stock model used in the recent past, BREHOMES, is validated against a second model, DECADE instead of validating against actual energy consumption. This thesis proposes to attempt to validate against actual electricity and natural gas consumption data recently collected by the UK Department for Energy and Climate Change for every census area at the Lower Layer Super Output Area level (LLSOA). Each LLSOA has around 500 households. Not every area will be suitable for validation as electric heating needs to be minimised in any area to more accurately validate the predicted non-heating end-use energy against the actual aggregated total electricity consumption. It

also requires knowledge of different billing regimes that encourage residents with electric heating to use a dual meter if they use electric storage heating.

The use of aggregated totals should be an advance in the understanding of the energy use of households and their appliances, lights, and electronics. However, there might be some resolution lost using this data, as ownership patterns of individual devices lose emphasis and the size of the household becomes more prominent as a predictor of this kind of energy use. In the opinion of this researcher, the accuracy gained from actual, instead of estimated, consumption, is worth the change in resolution.

1.4 Updating of the model using partial data

This thesis provides a secondary data analysis of major housing surveys and censuses that took place from 1996 to 2001 in England. As there are large gaps between these surveys and censuses that can be over a decade, plus additional years after the collection of the data for release to the research community, a way of updating the model using annually collected data can add value to the work. These surveys on energy use have only partial data regarding energy bills and not the full survey of meters over the course of multiple years, but they do paint a picture of the changing habits of energy use over the course of the interregnum between major surveys.

The use of this information is valuable for the researcher attached to individual household typologies. Modellers should be aware that using more easily obtainable aggregated figures would result in an ecological fallacy. Ecological fallacies are correlations made between sets of group averages instead of sets of individuals. As gaps between detailed surveys can be a decade or more, this information can help modellers take account of rapid changes in energy use behaviour.

1.5 Creation of a single-level model of household energy performance based on household size

In this thesis, a single-level model of non-heating end-use energy is developed using data from the fuel sub-sample of the 1996 English House Condition Survey enhanced by more recent social surveys up to 2008. The model is built using linear estimation methods using an interaction term combining the independent variables representing physical and human sizes of households. Finally, the model is validated against aggregated electricity use data of small areas collected from energy suppliers in 2008 together with census data from 2001. This leads to a simple model of energy use of appliances, lighting, electronics, and cooking in the existing housing stock in England.

It is not an ideal dataset to be using in the year 2011, but the fuel sub-sample of the 1996 English House Condition Survey (EHCS) currently is the most comprehensive and openly available database of actual energy use in housing. The survey, in addition to a comprehensive physical survey of the home and an extensive interview of the occupants, measured gas and electricity use in the home for nine consecutive quarters in the late 1990s between 1997 and 1999. This made it possible to exclude homes that used electric heating from this analysis. It is updated, however, with additional energy billing data from the Living Costs and Food Survey to estimate what similar homes surveyed in the EHCS might have been behaving like in 2008.

The survey data was used to create a linear model that predicted non-heating end-use energy using household size as an interaction term made up of the two independent variables of the number of occupants and the number of habitable rooms. The reasons for this are both historical and methodological in nature. The interaction term is the combination of physical household size and human household size. This combination has been historically been used in British modelling of energy performance in buildings as part of the Building Research Establishment Domestic Energy Model (Anderson et al., 1996). Methodologically, the energy use of households is positively skewed, which requires a transformation of the dependent variable of non-heating end-use energy to reduce bias in linear regression estimation techniques. The use of one interaction term instead of two independent variables enables the use of linear regression to represent the non-linear relationship between energy use and household size.

The model is then validated against the actual aggregated electricity use of LLSOAs in England for the year 2008 (Department for Energy and Climate Change, 2010b). These areas are same as those covered in the 2001 Census, with data covering the entire population on central heating ownership, physical household size, and the number of occupants. The areas considered are those with a high reported incidence of central heating. This minimises households with electricity as a primary heating source. Electric heating would obscure electricity used for non-heating energy in the aggregate data.

These steps result in a simple, single-level model of non-heating end use energy intended for domestic stock modelling. The building block of this bottom-up model, the energy performance of the household, does not need detailed building information. This is a deviation from SAP2009 which requires inputs for building geometry that could only be known through a site visit (BRE, 2010). Although there is a long and extensive research thread into the maximisation of the daylighting of homes due to the expense of artificial light before the advent of the electric light bulb (Tsao et al., 2010, BRE, 2010) and some cities have advanced LIDAR (light detection and ranging) mapping of the

built form, reconciling often complex arrangements of dwellings in urban areas with building outer surfaces in order to calculate daylighting in the existing stock is a difficult task. There are also differences between existing and new buildings in the use of low-energy light fixtures that have created a great deal of variation within similar-sized households. Although there have been energy-efficient compact fluorescent lamps (CFLs) available to replace fixtures intended for incandescent fixtures and are accounted for in the Standard Assessment Procedure (BRE, 2010), the rise of inefficient halogen fittings have added a third broad type of fixture to new housing. With the banning of the incandescent bulb and of the halogen lamp in the future, this variation may settle, but in the present housing stock, the presence or not of low energy light fixtures cannot be accounted for (Welz et al., 2011) and should be ignored when developing stock models of non-heating end-use energy.

The model is one that can be applied to the entire residential sector with confidence. The data is based on firm foundations in a national housing survey. It minimises the need for detailed, unavailable information required for correction factors for lighting use. Finally, it is verified against aggregated data for both the dependent and independent variables contained in the model.

1.6 Creation of a multi-level household energy performance model based on household size and type of area

The thesis moves on to examine whether group-level variance is present in addition to individual household-level variance when predicting non-heating end-use energy from household size. The hypothesis is one of homophily , or that “similarity breeds connection” (McPherson et al., 2001). However, the exact locations of these households remain unknown due to privacy rules. Area classifications are used as a viable alternative to actual location data of the individual dwelling and its exact relationship with its surroundings. Hierarchical linear modelling is employed to estimate group-level and individual-level effects for the same interaction term of household size on non-heating end-use energy (Steele, 2011). This results in a final model that estimates the group and individual level effects of household size on non-heating end-use energy.

Choosing to use more electricity for appliances, lighting, electronics and cooking should be explained not by just the homes that occupants live in, but the section of society that they inhabit as well as the external spatial influences around the home. The societal influences may be the ownership patterns of appliances and electronics being stronger is some areas than others. The spatial influences may put barriers to leisure outside of the home. It is not proposed in this work to

measure these influences, as direct causation is likely not to be proven. It is an output of this thesis to produce a number of plausible socioeconomic, spatial, and built environment causes of variation in non-heating end-use energy as a result of multilevel modelling for future investigation. In this work, the group-level effect is calculated by the mean household size and the individual-level effect the individual household size. This is done in order to maintain units of the effect size in kilowatt-hours of energy use instead of a unitless effect size that would have resulted from using different independent variables at the group and individual levels. The final output will describe the types of areas whose energy use can be effectively provided by a model of both individual and group-level effects using household size, and, more interestingly, describes the types of areas who cannot be adequately predicted by the model.

There is an opportunity that has recently been given to researchers to try to account for these connections through area classification information given to all the individual cases in housing surveys. The EHCS does not allow the exact location of any household to be released, meaning that the more nuanced relationships between the household and its surrounding urban design, retail and leisure offer, and overall transport accessibility cannot be assessed. However, the area classification method clusters spatial areas, in this case LLSOAs, where the differences between households are minimised and between areas are maximised using three pre-determined ranges for the final number of groups (<10, 20-30, 60+) based on 40 socioeconomic and built environment characteristics (Office for National Statistics, 2008b). Using multilevel modelling, the group-level effect of membership of an area type is shown to be almost as important to explain variance as the individual-level effect of a single household.

As this method splits the variance into group-level and individual household-level components, the group-level effect could be seen either as a “structural deficit” or “virtuous aid” to energy demand reduction. However, extreme caution should be used in making this interpretation. The group-level effects are based on the same interaction term at the group level, and not the many area-based characteristics that are not included in the modelling process. Instead, these characteristics are used as in descriptive mode as exact location data is not available for the individual households contained in the housing survey. Nevertheless, the multilevel analysis is a substantive contribution in the direction of using area-level, and not just individual-level data in assessing the non-heating end-use energy of the residential sector in England.

1.7 Thesis Structure

Following this introduction chapter, Chapter 2 describes in more depth the motivations for embarking on the project. The overall motivations for improving the modelling of non-heating end-use energy are described, including the fast growth of this kind of energy use in the home (2.2). A second motivation is the past emphasis on heating end-uses as they represent the majority of domestic energy use (2.3), including lack of verification of non-heating end-uses, and a consistent gap between predicted and observed energy use in current and past stock models (2.3.1-2.3.2). More emphasis on non-heating end-use energy in the context of England would be even more productive, as the UK, with a small incidence of electric heating, is a great laboratory to study these end uses with just electricity meter and billing data in a wide selection of homes and without the need to sub-meter end-uses in a more limited and invasive study of energy use in homes (2.3.3). Finally, the motivations for examining individual-level and neighbourhood-level data for predicting domestic energy use of the residential sector were examined (2.4). This analysis was based on new data sources on the area classification of individual households surveyed that became available during the course of the PhD, a new emphasis on forming a stock model from existing homes, and statistical modelling techniques available to deal with multilevel data (2.4.1-2.4.4).

Chapter 3 discusses the related work in the context of England. The definition of a domestic energy stock model is introduced (3.2) along with the historical context of energy crises and fuel poverty (3.3). The chapter then traces the history of the non-heating elements of the building performance model used in the UK called the Building Research Establishment (BRE) Domestic Energy Model in its first, second (3.4), third (3.5), and fourth (3.6) generations. This included the use of the building performance model to create a domestic energy stock model for non-heating end-use energy (3.5.5). The future directions for energy modelling are discussed (3.7), and the principles and methodology for area classification by the UK Office for National Statistics is introduced (3.8) leading to conclusions and potential research directions for this thesis (3.9).

Chapter 4 discusses the data sources required to satisfy domestic stock modelling of non-heating end-use energy. These include the requirements of models and stratification forming criteria for selection (4.2) and the many criticisms of the data of actual non-heating energy use that are available (4.3). The advantages and disadvantages of different sources of data for this project are discussed (4.4), including specialist housing surveys of households and non-specific housing surveys at the household and area levels (4.4.2-4.4.4). The chapter concludes by assessing the information available in the datasets assessed against the selection criteria, leading to a shortlist of datasets appropriate for this investigation (4.5).

Chapter 5 introduces the methodological specification for this thesis (5.2) along with two options for updating any domestic energy stock model for non-heating end-uses. One of these options allows updates around every ten years, and the other updates annually (5.3). A process of variable shortlisting and selection is then conducted for both of these options (5.4). The range of applicable quantitative methodologies for the model are explored (5.5), including top-down and bottom-up solutions as well as using area-level and individual-level data in single-level models (5.5.1-5.5.4) culminating in the exploration of using classed individual level and area level data in multilevel models (5.5.5).

Chapter 6 validates and prepares the data available to run statistical models which include housing surveys, expenditure surveys, and aggregate statistics. First, the datasets need to be reduced to just those homes that do not use electricity for space or water heating (6.2). The interaction term used in previous domestic energy models is retained with a key move from floorspace to rooms (6.3), and the nature of the reduced datasets as stratified or simple random samples is debated (6.4). As it is intended to use classical linear regression techniques to produce a single-level or a multilevel model, parametric tests are applied to the data, and transformation performed if necessary (6.5). Finally, the conversion process from the baseline data taken around 2001 to more recent years between major housing data collections is introduced (6.6). At the end of this chapter, the data is prepared and ready to be used in statistical modelling of the non-heating end-use energy of the residential sector in England.

Chapter 7 runs and validates two major models of the non-heating end-use energy of the residential sector in England – a single-level model and a multilevel model. First, the single-level model is run with the dependent variable of annual non-heating energy-use in a dwelling and an interaction term composed of the numbers of occupants and rooms (7.2). This is run using electricity meter data from a housing survey in the late 1990s and the 2001 Census data (7.2.1-7.2.2) and then updated to 2008 (7.2.3). The model then undergoes a validation procedure using aggregated electricity use data for selected census areas from the year 2008 (7.3). Second, a multilevel model is run with the same dependent variable, with the interaction term of households at the individual level and the mean interaction term at the census area level (7.4). This process includes the testing of different groupings (7.4.6), and verification (7.4.10).

Chapter 8 discusses the results and speculates on possible future work in domestic energy stock modelling of non-heating end-use energy at both the household and neighbourhood scale. First, the implications of retaining a single-level model are discussed (8.2) with a proposed change from floorspace to rooms as a measure of physical household size (8.2.2) to the problem of building a

bottom-up model from transformed data (8.2.4). The findings of the multi-level model are discussed (8.3), including the requirements of retaining group-level effect sizes measured in kilowatt-hours (8.3.2). Further possible domestic energy stock models for non-heating end-use energy are introduced (8.4), including hierarchical related regression, robust regression, archetypes, and conditional demand analysis (8.4.2-8.4.5). The discussion concludes by asking if domestic energy modelling is a pathfinder for the quantification of sustainable living (8.5), including the comparison of energy to less developed metrics of sustainability (8.5.1), knowledge transfer to non-energy aspects of sustainability (8.5.2), and what part the quantification of sustainable living has to play in the general assessment of the way communities are shaped (8.5.3).

Chapter 9 writes up the final conclusions of this thesis. The first main conclusion is that the multilevel model has real potential to accurately describe non-heating end-use energy in the residential sector in England (9.2). However, the interaction between group-level and individual-level effects needs further study with more targeted housing surveys by neighbourhood for further clarification. The second main conclusion is that the single-level model can be used immediately with annual household surveys to update the fuel sub-sample of house condition surveys that are taken a decade or more apart (9.3). However, the nature of the transformation of the data makes the inclusion of homes that are extremely large problematic. The final conclusion is that domestic stock modelling of non-heating end-use energy is a very reliable and stable way to measure progress towards environmentally sustainable neighbourhoods and lifestyles, and should undergo more extensive development with a wider range of statistical techniques in the future (9.4).

Chapter 2 - Background and Motivation

2.1 Introduction

The amount of energy used by non-heating end uses in the total housing stock in the United Kingdom has steadily increased, and at a considerably more rapid pace than the end uses of space heating and water heating. The purpose of this thesis is to examine ways of improving stock models of domestic energy consumption, and to question the routes taken to modelling in both academia and in building regulation. It is particularly interesting to examine electricity use within the context of the United Kingdom because of its unusually complete separation of fuel types into natural gas for heating and electricity for non-heating end uses. This thesis hypothesises that a synthesis of individual household-level and area-level predictors can provide the basis for significantly improved predictions of non-heating end uses of energy, defined as appliances, lighting, electronics, and, when fuelled by electricity, cooking. The author's expectation is that this could shed new light on the ways that energy efficiency measures and proposed and funded in the United Kingdom.

Finding the right tools and predictors of to create an accurate estimate of non-heating end-use energy for the residential sector by using a bottom-up stock model has proven to be difficult. The UK government made a choice to pursue a bottom-up instead of a top-down model in order to relate single dwelling energy models for building permission with the housing stock model (Kavgic et al., 2010). Instead of driving down actual energy use in buildings, conclusions have been drawn recently that energy modelling has only encouraged improvement in the energy efficiency of modelled buildings (Day et al., 2007). Historically, energy use in dwellings when occupied has been greater than the modelled energy use (Lowe et al., 2009, Zero Carbon Hub, 2011, Lorimer, 2010). This thesis argues that this problem has reached a salient point for researchers and practitioners at the time of writing regarding the difference between energy consumption as planned, as built, and as occupied.

A second key research question that is also important is the treatment of individual-level, household-level, and area-level characteristics that predict electricity use for non-heating end uses. Targeting households for energy-saving measures for the home in the United Kingdom are devised on the basis of either information derived from the individual home or from the aggregate totals of a group of buildings that are measured by area. The availability of high-quality data on the full spectrum individual households, their dwellings, and energy use is limited by data protection rules

that are analogous to many worldwide. The approaches for estimating these quantities are very different statistically and the differences between them will be explored, explained, and compared using statistical methods that are applied to parametric data drawn from households, and aggregate data drawn from small areas throughout England.

Lastly, some of these techniques can aid the makers of policy for the built environment to decide whether energy efficiency measures or on-site renewable electricity generation methods, applied to a general area can be measured using only area-level data that contains the summary of characteristics of all households. This technique, called ecological inference or neighbourhood modelling, is statistically inadequate compared to techniques that use household-level data, but should be considered under two conditions. The first condition is that data protection rights include household energy use in the foreseeable future. The second condition is that the researcher must assume that the propensity of a socioeconomic grouping of households to use energy does not change from area to area. Without these two conditions, any results from ecological inference would be unsound because violating the former would mean that data at the quality level of the aggregate is available at the individual household-level and should be used, and violating the latter would introduce a number of confounding factors that would halt any meaningful analysis of neighbourhood-level data. This thesis will argue that in the context of energy, buildings, and cities, these conditions are reasonable.

2.2 Motivation for improving modelling of non-heating end uses

2.2.1 Introduction and terms of reference

This thesis will specifically focus on the modelling of non-heating end-use energy in households. Modelling is the use of data to predict an outcome. The outcome in this thesis is non-heating end-use energy. “Energy use” will be used in this thesis as a short-hand for “end-use energy”, or the energy delivered to a user within a building (Sørensen, 2011). An end-use is the range of ways that energy can be used by people within the context of a home – for example, a user can turn on (or set a timer for) heating, or turn on a light. A non-heating end-use for homes in the United Kingdom is defined by this author as energy used for appliances, lighting, electronics, and cooking and not for space heating and cooling or water heating. The unit of measurement of energy will generally be kilowatt-hours per annum to account for significant seasonal variation in energy use.

The UK National Statistics harmonised survey definition of a household is “one person or a group of people who have the accommodation as their only or main residence and (for a group) either share at least one meal a day, or share the living accommodation, that is, a living room or sitting room”

(Office for National Statistics, 2008a). It is often, but not always, interchangeable with the definition of a residential dwelling as self-contained accommodation, and only sometimes interchangeable with the definition of a residential building, especially in urban areas (Census Advisory Working Group, 1999).

The scope and nature of modelling of energy use in housing can be split into two distinct, but related forms of modelling: building performance simulation and housing stock modelling. Building performance simulation is a computer model that predicts the energy demand of an individual dwelling when it is used in real life. Housing stock modelling predicts the energy demand of the residential sector of a given entity. Typically this entity is a nation or region. The results of a housing stock model are used to inform policy choices at the national level from the macro-level issues of generation to the micro-level issues of choice of technology to promote for implementation within buildings. A lack of understanding of the relationship between building performance simulation and housing stock modelling has led to a misappropriation of terms of reference in the planning and built environment professions not directly associated with environmental engineering (Mayor of London, 2009, Homes and Communities Agency, 2007).

Modelling end-use energy outside of heating is of interest because

- it continues to grow much faster than energy use overall,
- most research has been done on space and water heating,
- of the historical gap between predicted and actual energy use in existing models,
- of the lack of documentation on models in building regulation, notably in the UK,
- because of the unique characteristics of the UK housing stock that enable the measurement of these end-uses with electricity meters from general housing surveys and aggregate data,
- and of the lack of clarity from the academic community on whether socio-technical, statistical, or engineering approaches to non-heating end-uses are appropriate.

This section will consider the first five motivations that are non-methodological in nature.

2.2.2 Non-heating end-uses are growing faster than heating end uses

Data from the 2009 Domestic Energy Fact File for the United Kingdom in Figure 2.1 reveals that from 1970 to 2006, the energy consumption of lighting and appliances grew by 148 per cent compared to 23 per cent in overall energy use (Shorrock and Utley, 2008). This is supported by other data released by the Office of National Statistics in the UK (Office for National Statistics, 2010).

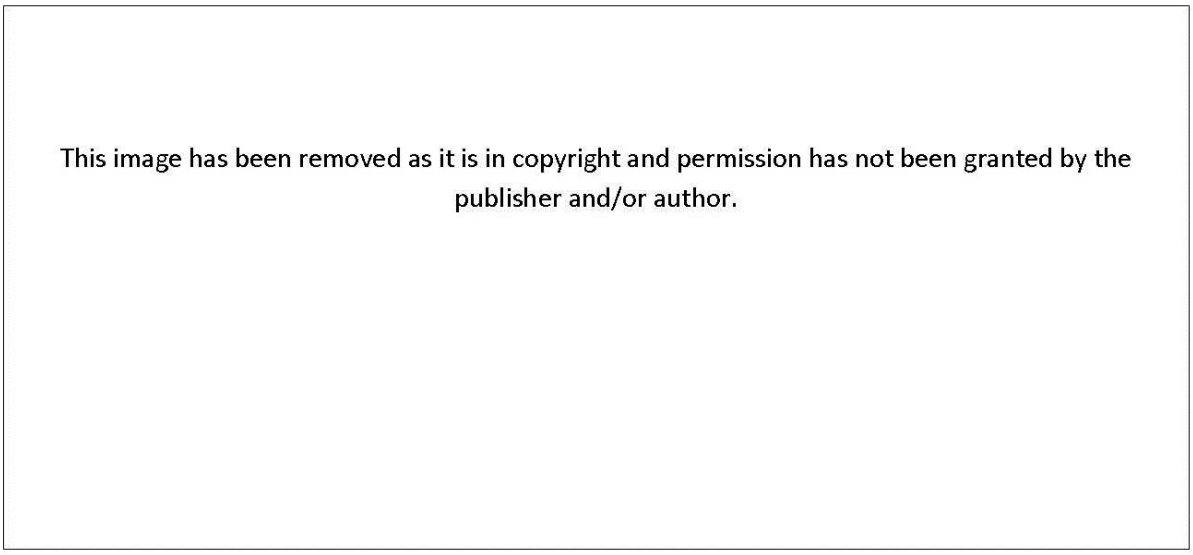
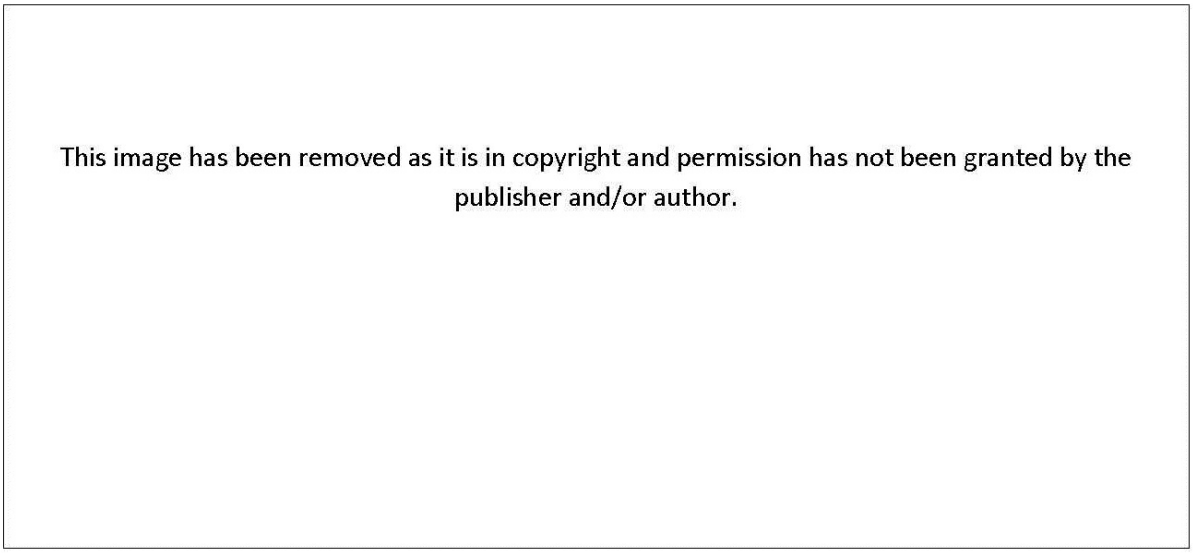


Figure 2.1: Domestic energy consumption by end use, 1990 to 2006 (adapted from Shorrock and Utley 2008)

The main driver of domestic energy consumption in this sector in the UK has been information and communications technology (ICT) and consumer electronics (CE), outbalancing falls in lights and "white goods" (Energy Saving Trust, 2011, Department for Environment Food and Rural Affairs, 2009). The effect of appliance efficiency is contested. One example, presented by the Energy Saving Trust, predicts a fall in energy consumption in lighting and refrigeration due to efficiency improvements and a modest increase in ICT and CE between 2009 and 2020 as show in Table 2.1. In contrast, the International Energy Agency (IEA) predicts bigger increases of around 50% more units and 25% more consumption per unit in the same time period as shown in Figure 2.2.

Table 2.1: Past, current, and future electricity consumption in the UK residential sector (Energy Saving Trust, 2011)



There has been, and will continue to be advances in the energy efficiency of lighting and refrigeration with the phasing out of incandescent and halogen light bulbs and as a fridge built in the mid-1990s consumes on average about 50% more electricity than a comparable model today. However, in the period 1990 – 2030, the IEA claims that the average unit consumption of electricity for these uses in Organisation for Economic Co-operation and Development (OECD) countries will have risen by 75 per cent – almost 2 percent per year on average (International Energy Agency, 2009). Clearly, although the climate change impact caused by lights, appliances, and electronics is less than for heating and there are plans to decarbonise electricity generation (Department of Energy and Climate Change, 2009b), the growth in electricity demand may well be significant and needs further investigation. This work will seek out alternative sources in order to understand the non-heating end-uses of the residential dwelling.

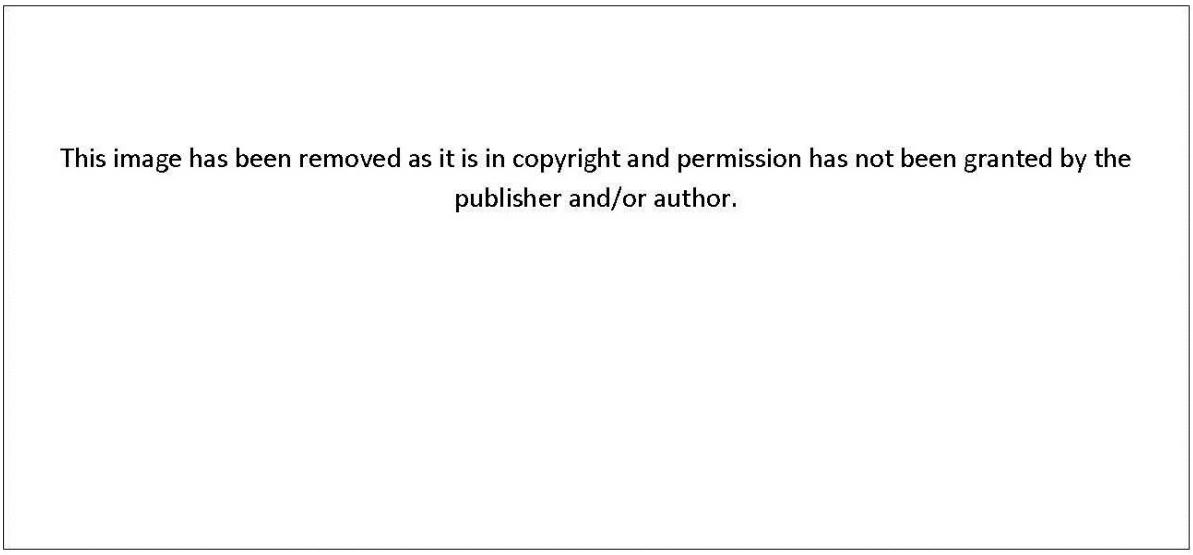


Figure 2.2: Estimated change in stocks and average unit energy consumption of residential ICT and CE appliances in the OECD, 1990 to 2030 (International Energy Agency 2009)

Data recently released in the United States, which has a similar growth in lights, appliances, and electronics energy consumption to the United Kingdom, reveals a striking trend in the growth in these sectors from 1978 to 2009. Personal computers and televisions can be highlighted as examples. In 1978, personal computers were rarely seen in homes, but by 2009, 76 percent of U.S. homes had at least one computer, eight percent higher than 2005, and 35 percent had more than one computer. In 1978, most households only had one television. By 2009, the average household contained, on average, 2.5 televisions. Furthermore, screen size has grown, which is correlated with average energy consumption per television, with almost half of all homes containing a television with a 37 inch screen or greater. The most striking trend is that the same proportion of the population has at least four rechargeable electronic devices (U.S. Energy Information Administration, 2011).

2.3 Past emphasis on research on heating end-uses within regulation and academia

The majority of research on building performance models has been done on improving the efficiency of space and water heating since the advent of the discipline following the energy crises of the 1970s. The response to this crisis was active involvement by the government in directing research in energy use in the built environment. The Building Research Establishment was formed in 1970 in a merger of previously independent government research centres (such as the Building Research Station) dealing with the built environment, its methods, materials, and threats – with heating the

primary target as it constituted 86 per cent of all energy consumed by the domestic sector (Office for National Statistics, 2010).

Work in support of building regulations in place in the United Kingdom became a central feature in the BRE's work. Energy and environmental studies, stimulated by the energy crises in the 1970s, came into the BRE in the form of energy conservation measures in housing and other buildings. This work received new emphasis in the next decade, with a large programme promoting energy efficiency in buildings comprising 70 staff in 1993 (Courtney, 1997).

One result of this work was the development of a building simulation model called the BRE Domestic Energy Model (BREDEM), last revised in 2009, with the supermajority of its algorithm addressing the modelling of space heating. BREDEM can be described as a building physics algorithm with variables such as boiler type, insulation, and air permeability taking prominence in explaining the variance between homes over housing size in terms of floorspace or numbers of occupants demanding heat. As an illustration, Figure 2.3 below shows the amount of pages devoted to each end-use for the BREDEM-8 model in use from the mid-1990s. The core of the BREDEM model from its beginnings was:

$$E_d = \frac{(\sum AU + c_v)DD}{f} \quad (1)$$

Where E_d is the delivered energy for space heating (kWh), $\sum AU$ is the area weighted sum of U-values of all external surfaces (W/K), c_v is heat lost through ventilation (W/K), DD is the length of the heating season (days), and f is the efficiency of the heating system, defined as the ratio of the heat delivered into the dwelling to the calorific content of the fuel (Uglow, 1982).

BREDEM has for some years been the predominant model in use throughout the United Kingdom. For the purposes of this work, the scope is limited to building regulations that apply to England only due to planning and building regulation being a devolved power in the United Kingdom. The version of BREDEM published in the public domain and approved by the English government for use in building control and the implementation of the European Directive on the Energy Performance of Buildings is called the Standard Assessment Procedure (SAP) (BRE, 2010, Department for Communities and Local Government, 2008b).

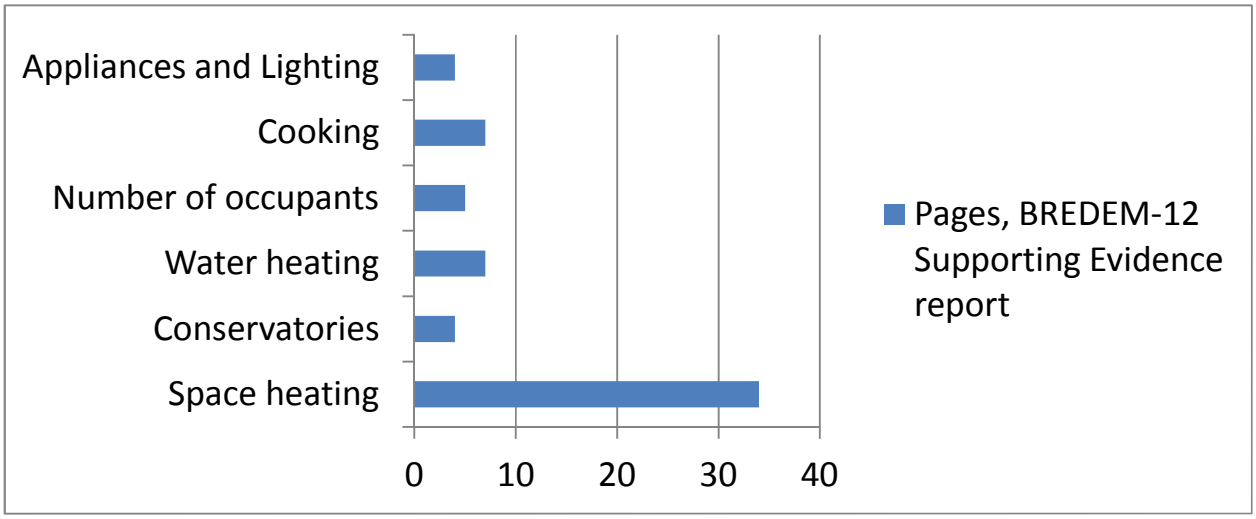


Figure 2.3: Number of pages in the supporting evidence report for the 2005 version of the Standard Assessment Procedure in the United Kingdom (Energy Advisory Services 1996)

Academic research in energy use in residential buildings has been limited by availability of detailed data in the United Kingdom. The prevailing approach in academia for heating end-uses has been to apply engineering and technical knowledge methods (Chapman et al., 1985a, Lowe et al., 2009). This considers a residential dwelling as a collection of devices or materials whose performance can be optimised (BRE, 2009). Therefore, work on building performance has informed the modelling of an individual dwelling, and estimates for heating end-uses in the residential sector can be made by summing up individual dwellings. Chapter 3 contains details of historic academic work in the energy performance of dwellings.

In non-heating end-uses, the number of devices or appliances is less clear at the individual dwelling level. The requirements for measuring individual appliance usage are onerous in terms of cost and manpower, and studies that sub-meter electricity use into its components are extremely rare and have had low numbers of participants sampled to represent a targeted segment of the population instead of a national population (International Energy Agency, 2007, Macmillian and Kohler, 2004, EURECO, 2002). The response of this thesis is to focus first on housing surveys, then on aggregated energy use for small areas that offer less resolution than sub-metering but are collected with methodologies that attempt to include all segments of a national population of households and residential dwellings.

2.3.1 Lack of detailed documentation on verification of building performance models used in building regulation of non-heating end-uses

The BRE was part of the Department of the Environment and its research became focussed on government policy and regulation following the adoption of recommendations of the 'Rothschild' report *A Framework for Government Research and Development* that placed more research money

into government-sponsored corporations instead of universities (Command 4514, 1971). This focus on supporting policy with research was a consequence of the contrasting approach to academics in the built environment. Academia in the United Kingdom was still funded according to the Haldane principle - that research grant budgets are placed into research councils that are free from political and administrative influences (Haldane, 1918). This partially explains parallel streams and divergence of approaches to energy use modelling in the built environment.

Work in support of the building regulations became a central feature in the BRE's work. Energy and environmental studies, stimulated by the energy crises in the 1970s, came into the BRE in the form of the writing of energy conservation regulations for buildings. In 1997, the BRE was privatised. Government work that previously was given to the BRE for research and dissemination was now put to tender with the group needing to bid for this work against other consultancies, leading to a decrease in dissemination and invitations for academics and practitioners to scrutinise and critique this work now that the BRE's intellectual property was commercially valuable and no longer government-sponsored (Courtney, 1997).

The documentation presented by the BRE on the formulation of the Standard Assessment Procedure (SAP) has not been open to outside scrutiny since the 1996 version. The BREDEM model which is published as SAP is verified against a simulation of the entire UK housing stock using a stock model called BREHOMES. According to interviews conducted by the Policy Studies Institute (Ekins and Dresner, 2006), BREDEM calculations are taken for typical homes in a number of categories of dwellings as derived from national housing surveys, both public and proprietary, then multiplied by the total number of homes in each housing category. For appliances and lighting, the DECADE model (Environmental Change Institute, 1995) simultaneously creates another bottom-up model of domestic energy use for all electronics and appliances that are owned in the UK along with their usage patterns and power consumption. The BREDEM calculations are changed by a scaling factor to match the DECADE estimate. They are then compared with a top-down estimate from the *Digest of UK Energy Statistics* (DUKES). The BREDEM calculation is adjusted again by altering the assumed temperature demanded by occupants – this is all called a 'reconciliation procedure'. (Ekins and Dresner, 2006). Chapter 3 will provide a longer version of the history of BREDEM and the data that supports its algorithm.

There have been a few attempts in academia to create a rival modelling method. One example of such as undertaking is the UK Domestic Carbon Model (UKDCM) – now in its second version - which was specifically built to predict the measures need in 2050 to reduce carbon emissions by 40% from 1990 levels (Boardman, 2005, Hinnells et al., 2007). This was a model built out of demographic data

from the national housing surveys that were publicly available. This is still a bottom-up model built from individual dwellings. UKDCM predicts the expected energy consumption of a dwelling to get to a level of heating and use of lighting and appliances as determined by its group. The entire building stock is disaggregated by tenure, dwelling type, dwelling age, construction method and English region. Again, the DECADE model is used to estimate the total amount of demand for appliances and lighting. The characteristics of heating demand in homes in 1996 were adjusted over time to 2050. UKDCM also "built" new homes that would exist in 2050 to satisfy new household formation and assumed a replacement rate. The UKDCM is not available for researchers outside of the Environmental Change Institute, nor was it intended for use on individual dwellings or in the present day housing stock.

This thesis will take a different approach. Instead of verifying a housing stock model of totalled individual building performance against the national energy use of electricity, it will attempt to verify it using aggregate totals from areas across England. The 'reconciliation process' for heating in BREHOMES was by a change in assumed internal temperature in dwellings and therefore related only to heating. The DECADE model is a model of the national ownership of goods. A new process based on geographic areas is intended to detect latent effects that are attached to areas or a classification of area. It is hypothesised that these correlate with either socioeconomic groupings or typologies of the built environment.

2.3.2 Gap between predicted and actual energy use for non-heating end uses

There is one large issue with the models developed to date is that there is a gap between the modelled prediction and the actual energy use (Summerfield et al., 2006). In more recent research project run as a partnership between architects and built environment researchers, the data demonstrated that real building energy use significantly exceeds design expectations (RIBA / CIBSE, 2009). Post-occupancy evaluation projects in the United Kingdom focus on the non-domestic sector and therefore studies on housing are limited (Bordass, 2007, British Council for Offices, 2007, Bechtel, 1980).

The gap between the predicted and modelled energy use is well-documented in the area of heating demand in the context of the United Kingdom (Ekins and Dresner, 2006, Sorrell et al., 2009, Wingfield et al., 2008). Academics in the field of building performance have also observed this gap in the measurement of energy savings after interventions that decreased the modelled energy demand for heating (Energy Saving Trust, 2004, Hong et al., 2006b).

The same research has not been replicated within academia for non-heating end-uses, but there has been a significant revision upwards inside the BREDEM model for these end-uses from 2005 to 2009

after criticism in a government-commissioned peer review of the methodology, interestingly, due to its impact on the estimate of households in fuel poverty (BRE, 2009, Sefton and Chesshire, 2005). The result of this underestimate was that the amount of energy consumed by the UK building stock for these end-uses from 2005 to 2010 under SAP2005 was estimated to be around half the reported energy use derived from aggregate totals for small areas (Lorimer, 2010). This thesis will examine the extent to which this gap has been closed and the amount of variance that is explained with the predictor variable of usable floor space used in both the 2005 and 2009 versions of SAP.

2.3.3 Use of electricity primarily for non-heating (and non-cooling) end-uses in the United Kingdom

In most of the developed world, research in energy use is hampered by the lack of convergence of the type of fuel consumption and energy end-use. The United Kingdom, however, is a valuable place to investigate non-heating end-uses because a supermajority of homes have central heating, and use natural gas instead of electricity for heating fuel (Fawcett, 2000, U.S. Energy Information Administration, 2011, Shorrock and Utley, 2008). In addition, central heating systems that use natural gas in most cases use the same heating system and fuel for space heating and water heating (Peter Warm, 1999). It is typical of a northern European nation to have a low electricity load for cooling in the residential sector. Therefore, when examining the residential consumption of groups of buildings in the UK, electricity consumption can, if there are a high percentage of buildings in that group report using central heating, be equated with non-heating end-uses.

Figure 2.4 illustrates examples of how space heating in the developed world is delivered using a variety of different methods depending on domestic fuel sources and imports, climate, and scale of the heating system. In the United Kingdom, three forces combined to make natural gas the fuel of choice for space heating: the discovery of natural gas in large quantities in the North Sea since the 1970s, the lack of demand for cooling, and cultural choice in having one heating system per household. In the United States and France, this resource of large past supplies of natural gas in relation to demand did not existing, cooling demand is much higher in the southern parts of the country which drives electric dual heating and cooling systems. In some parts of the United States and Scandinavia, hydroelectric power is viewed as an abundant resource which drives takeup of electric heating. It should be noted that although cooling can require significant amounts of electricity in the future, and cooling is newly added into SAP in 2010 (BRE, 2010) the time period covered in this study does currently not have evidence of large-scale residential, as opposed to office and commercial, cooling in the UK (Adnot, 2003). The combined effect of these factors creates a mix

of heating and cooling fuels in most developed countries in the world apart from the United Kingdom.

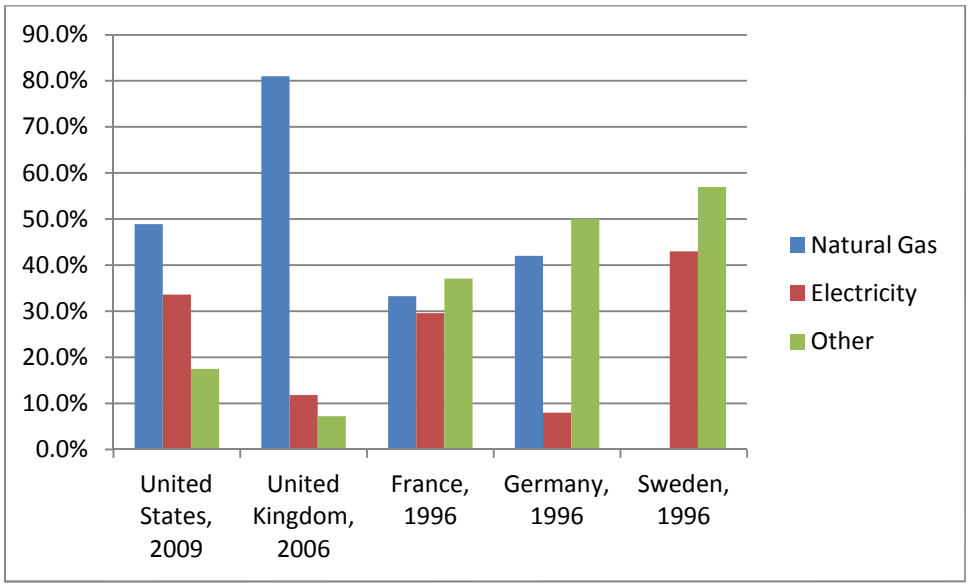


Figure 2.4: Space heating fuel types present in residential dwellings by country in representative developed nations (Shorrock and Utley, 2008, U.S. Energy Information Administration, 2011, Fawcett, 2000)

This leads to the discovery that the UK has a preponderance of natural gas-fed, single household heating systems. For the researcher, this creates a significant benefit which allows the removal of electric space heating use from large numbers of homes as a confounder when correlating electricity consumption with non-heating end-uses. Energy fuels in the home in the United Kingdom have continuously diverged to natural gas as the fuel for space and water heating end-uses and electricity for all other end-uses (with cooking split between the two, representing less than 4% of household energy demand) since 1970. Statistics from the *Digest of UK Energy Statistics* state that centrally-heated dwellings have risen to about 90 percent of the total housing stock, with about 90 percent of central heating systems using natural gas as fuel (Shorrock and Utley, 2008).

This thesis will take advantage of the unique nature of the relatively low uncertainties of equating electricity consumption to non-heating end-uses in the residential sector in the United Kingdom particularly when using data disaggregated to the LLSOA level. It will also restrict itself to the context of the UK's housing stock instead of attempting to make its conclusions internationally applicable. The isolation of these end-uses without depending on small studies of sub-metered dwelling gives the researcher larger sample sizes from general housing surveys and analysis of area-level effects where aggregate information is available.

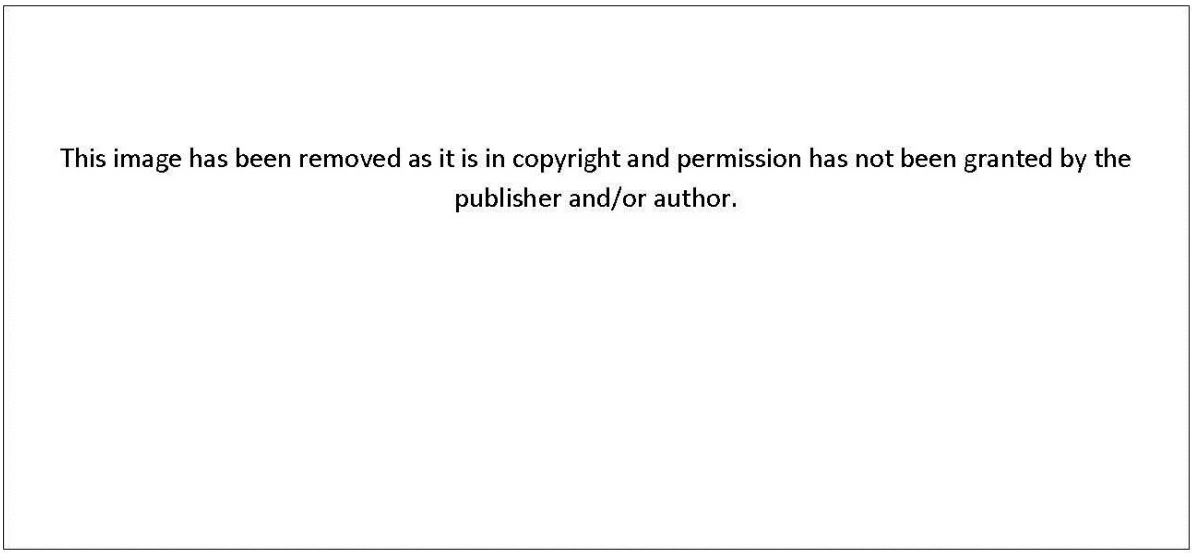


Figure 2.5: Centrally heating dwellings as a percentage of all homes – 1970 to 2006 (Shorrock and Utley 2008)

The United Kingdom is also unique in the provision of water heating as an ancillary to the central heating system. General housing surveys , including the English House Condition Survey and other specialist surveys, reveal that amongst homes that possess central heating, the use of natural gas for water heating is historically reported at slightly under the same rate as for space heating (88 per cent as opposed to 90 percent) (Peter Warm, 1999, Environmental Change Institute, 1995). Therefore, use of electricity for hot water heating is no different from space heating as a confounder that might mask non-heating end-uses.

However, cooking consumption represents a mix of electric and/or gas hobs (stovetops) and ovens. National trends in ownership and usage of hobs and ovens can be merged to create a proportion of electric-to-gas cooking consumption – this is trending towards electricity and away from natural gas (Market Transformation Programme, 2008). Cooking as an end-use has declined steeply in the last 30 years to a level of only four percent of all kilowatt-hours in 2006 (Office for National Statistics, 2010, Shorrock and Utley, 2008). Cooking using electricity as the fuel is a confounder that needs careful consideration as a non-heating end-use as only part of this use can be assigned to electricity consumption in dwellings, and there is little data below the national level on ownership levels of gas and electric cooking appliances.

2.3.4 Conclusions

Modelling the energy end-uses outside of heating is of interest to the academic community because of the opportunities for studying effects outside of the current convention of household size as measured by usable floor space. This is because

- the same amount of floorspace is predicted to use more non-heating end-use energy because of rapidly growing ownership patterns of electronics while heating demand remains stable,
- all segments of the residential population of the UK are available for study in both aggregated and disaggregated forms,
- the gap between predicted and observed energy use is still poorly understood,
- the lack of publicly available methodologies is an issue in evaluating regulatory models of energy use by academia ,
- and the UK is a very good laboratory for research in this area because of the relatively high correlation between delivered electricity consumption and non-heating end-use energy.

These effects will be studied from a social science perspective and the extent to which stock models of non-heating energy use can be studied as both a physical construct and as an urban system.

2.4 Motivation for examining the question of household-level or neighbourhood-level data

2.4.1 Introduction

This thesis is intended to develop a modelling and validation method for simulating the non-heating end-use energy of a household using both individual household and area level predictors. This simulation is of a household, and not a dwelling or residential building. Therefore it is similar to building performance simulation models for energy use, but is assessed on the household, and not on the dwelling. For heating end-uses, consideration of thermodynamic properties of the dwelling and its building (if part of a multi-unit complex) would prevent this approach. Therefore, it is an approach that is tailored for non-heating end-uses taking a statistical approach based on multilevel modelling and ecological inference. This approach does not use socio-technical or engineering methods to total up all end-use devices. This would require the researcher to estimate the number of devices in the household that demand electricity and sum the product of each device's power demand rate with load factor; instead, it uses the characteristics of the household and its area as predictors.

In order to conduct this research, quantitative analysis was conducted on predictor and outcome variables at both the individual and the neighbourhood level. This is possible within the context of England because of the wealth of data at both the household and neighbourhood levels that are designed at the national (England) level. This data, however, has been interpreted by regulators and academics from socio-technical and engineering perspectives, but in the opinion of this author, not

as effectively from a social science perspective. Furthermore, the modelling techniques employed can produce large Type I “false positive” or Type II “false negative” errors if they use incorrect assumptions, inappropriately assess outliers, or improperly re-scale common outcome variables of consumption of non-heating end-use energy. However, all approaches thus far only use variables at the household level or below, and no area-level effects are tested and therefore unknown. This thesis will establish a statistical modelling basis for investigating any area-level effects that are significant in England because:

- housing surveys are available that can isolate non-heating end-use energy in individual households with limited location data,
- social science quantitative techniques are available to study non-heating end-use energy of individual households within socioeconomic and built environment groups, and
- aggregated electricity consumption data that covers the entire population and is widely available and frequently updated.

2.4.2 Data sources

Measurement of non-heating energy end-uses in England at the national level can be done with three types of dataset: housing surveys, device ownership trends, and small-area aggregate statistics. There is one recent housing survey that asked participants to record metered electricity data along with recording the heating fuel of the household – the 1996 English House Condition Survey (Department of the Environment Transport and the Regions, 2000b). Ownership patterns of devices that demand electricity are described throughout numerous propriety consumer surveys, but for energy use in the household the main metrics of ownership trends are contained within the Market Transformation Programme research centre (Market Transformation Programme, 2010). Detailed information on the makeup of the built environment throughout the United Kingdom is contained within the releases of the 2001 census into aggregated sections because individual entries are not available due to the Data Protection Act (Carey, 2009), and the amount of electricity use of these areas is covered by metered data collected by the Department for Energy and Climate Change (Department for Energy and Climate Change, 2010b). These are Lower Layer Super Output Areas (LLSOAs) – for simplicity, the word ‘neighbourhood’ in the remainder of this thesis is a colloquial shorthand for an LLSOA. These data sources have different strengths and weaknesses for social science research. Chapter 4 will contain detailed information on the collection methods, types of variables, and distributions of variables in all of the datasets.

The source of individual household data on energy use at the national level is through housing surveys. Some of these surveys are conducted by the public sector and therefore the data is open –

the English Housing Survey is the main current survey (Department for Communities and Local Government, 2010a). There are also commissioned surveys which are not available to academia, notably the surveys commissioned by the Building Research Estabishment (BRE) to support the development of the BREDEM algorithm (Impetus Consulting Ltd, 2006). Unfortunately for the energy research community, the last survey with open data was the 1996 English House Condition Survey (EHCS). A fuel sub-sample was taken from this survey where participants kept a 27-month diary of their gas and electricity meters. Combined with household and occupancy characteristics, this dataset is the richest and most complete picture of energy use in English households to date (Department of the Environment Transport and the Regions, 2000b). However, the data is becoming out of date, and there are other surveys, notably the Living Costs and Food Survey (LCFS) that are available that ask respondents about their last household energy bills with additional household data that can enable an updated estimate of energy use in the present day when harmonised with the 1996 EHCS (Office for National Statistics and Department for Environment Food and Rural Affairs, 2010).

The Department of Energy and Climate Change (DECC) have released data on energy consumption in both the domestic and non-domestic sectors delivered to small areas since 2004. These areas, called Lower Layer Super Output Areas (LLSOAs), were first introduced in the 2001 census as a new statistical standard. There are several advantages to using LLSOAs as a basis for energy use statistics; they are relatively consistent in terms of population (minimum population of 1,200 equating to around 500 households, and numerous possible socioeconomic and built environment predictor variables can be drawn from census data. Data in DECC's small area statistics database is shown by consumption in kilowatt-hours split by ordinary residential electricity, economy7 residential electricity, industrial/commercial electricity, domestic gas and industrial/commercial gas, number of gas and electric meters and average consumption per meter. The diversion of heating energy end uses to natural gas and lights and appliances to electricity as fuel in the home means that this energy data is useful for assessing the non-heating end-use energy. In some cases, heating is supplied by electricity, but there is census data for each LLSOA to warn the researcher of the prevalence of this confounding factor. (Department for Energy and Climate Change, 2010b).

The main weakness of the data available for the energy researcher are first, the number of households surveyed, and the lack of repeat surveys with the same energy composition diaries that enable longitudinal analysis. The housing fuel sub-sample of the 1996 EHCS of 2,531 records and the 2008 LCFS total of 7,000 households is very small compared to other areas covered by social science quantitative analysis methods. Examples in each fields with a national (England) survey include

health (13,000), citizenship (15,000), and crime, with an astounding 66,000 records collected each year (Economic and Social Data Service, 2011). This lack of data is compounded by not performing repeat measurements. If the survey can continued annual in the same form, almost 33,000 records of household energy use would have been collected and our understanding could be correspondingly greater. In comparison, there are 32,482 LLOSAs with yearly mean end-use energy consumption data. Therefore yearly energy data collected at the aggregrate level, despite the values representing the mean of a collection of buildings, should be viewed as valuable by the researcher in a way that researchers in other areas do not and should not.

2.4.3 Approaches to building simulation and stock modelling, with household simulation as an alternative for non-heating end-use energy

This section will review the differences between building simulation and stock modelling, and introduce the concept of household simulation as an alternative methodology for non-heating end-uses energy modelling. Building simulation modelling in the UK has been separated into two pieces: a thermodynamic flux model for heating, and a mix of engineering and socio-technical (sometimes called techno-economic) modelling techniques for non-heating (BRE, 2010). Housing stock modelling contains a larger and more diverse range of methods, from top-down econometrics to bottom-up statistical and engineering approaches. Housing stock modelling has been extensively used as a validity check and feedback mechanism for building simulation models in the UK and internationally (Ekins and Dresner, 2006, Swan et al., 2008, Shorrock and Dunster, 1997).

The current literature and statistics regarding energy modelling and energy use by occupants of their appliances and lights in domestic buildings show that simulation of building performance has been approached using mainly sociotechnical and engineering approaches, and not by social science approaches that can accommodate analysis of group membership of households. The equations generated by building performance simulations then are aggregated into the energy performance model of the residential sector of an area through a housing stock model using a variety of econometric, statistical, socio-technical, and engineering techniques.

Building simulation modelling is defined as the estimation of all end-use energy requirements of a residential dwelling. Heating can be provided by either end-use energy – in other words, a heating system - or by other means such as passive heat gains. Some passive gains come from appliances, particularly cooking, to the heating of its area of the building, known as a heating zone (Shorrock, 2010). A second well-studied passive heat gain is solar gain through windows during the daytime (Doran and Anderson, 1995, Thomas and Fordham, 2003). Non-heating end-uses do not contain such confounders from heating end-use energy, therefore they can be treated as a separate system of

energy consumption in the home. The collection method for simulation is by engineering archetype - by testing a limited set of dwellings that represent the groupings of homes that are representative (Parekh, 2005). Building research centres can provide a set of representative homes for testing – examples are the Milton Keynes Energy Park and the BRE Innovation Park (UK Energy Research Centre, 2008, Gaze, 2008).

Building simulation of energy consumption is largely an engineering exercise in predicting the heat balance of a new building as designed. Building design tools which predict energy use from a designed building, such as TAS, ESP-r, and EnergyPlus, depend on algorithms of energy consumption developed by building science laboratories. Two leading examples of theses algorithms are BREDEM, developed by the BRE in the UK context, and DOE-2.1E with IBLAST, developed by the Lawrence Berkeley National Laboratory in the United States (Crawley et al., 2008). As the heat gain through the operation of active heating combined with passive gains is the focus of these algorithms, research on non-heating end-uses has typically centred on the question of the rate of passive gains caused by this type of electricity consumption.

These should not be confused with green building rating systems such as LEED, BREEAM, the Code for Sustainable Homes, and SBTool. These are systems that allocate points for energy use, site selection, transport use, biodiversity, water use, and other categories defined by the awarding body as included in sustainability. These definitions are shaped by the rewarding body that developes the tool. An award from any of these green building bodies does not necessarily indicate exemplary energy performance of the building (Sedlacek and Maier, 2010).

In building performance models, there are detailed estimates of the energy consumption of low-energy lighting and lowering demand through daylighting as a by-product of testing their impact on the heat balance. However, the remaining non-heating end-use energy modelling is derived from housing survey information, therefore non-heating end-use energy estimation becomes a hybrid of engineering archetype and socio-technical techniques (Sefton and Chesshire, 2005). Socio-technical techniques are those that estimate the amount of interaction humans have with technology and develop an estimate of energy use for each devise based on survey data in support of each appliance. However, this kind of data is rarely collected as sub-metered energy use in households; instead, consumption is estimated with household size. Further discussion on the history of the development of a mix of techniques, notably in the development of BREDEM in the UK context, can be found in Chapter 3.

Reviewers of housing stock models have described the methods for analysis of the residential sector as top-down and bottom-up (Swan and Urgusal, 2009, Kavgic et al., 2010). Top-down methods estimate energy consumption of the residential sector using national figures and treat the sector as what Swan and Urgusal call an “energy sink”. Bottom-up methods sample the energy consumption of end-uses, individual dwellings, or groups of homes that are representative of the real population of dwellings contained within a region or nation.

However, the line between the building simulation model and the housing stock model becomes increasingly blurred in the process of checking the validity of building simulation models. Housing simulation equations are extensively used in the assembling of building stock models in the UK as a way of analysing the housing stock of an area or region. However, there is variance between the sum of all dwellings estimated using a building simulations model from the reported aggregated consumption of a nation or region. This discrepancy can emerge if the data, methods and approaches come from different sources. Housing stock models claim to have greater statistical power than building simulation models because they are derived from housing surveys or from official aggregated statistics than encompasses the entire population, and therefore building simulation models are altered to decrease this variance. In the UK, the BREDEM model was altered by changing the assumed internal temperature of a typical dwelling (Shorrock et al., 2005, Natarajan and Levermore, 2007, Kohler and Hassler, 2002). In Chapter 3, the process of reconciling this variance is explained.

This thesis proposes to break away from building simulation to a slightly different practice of household simulation for non-heating end-use energy. Household simulation uses housing surveys and aggregated data instead of archetypes to establish its initial estimates, and therefore the variance is due to the uncertainty in estimation, not from data sources or hybridisation of techniques. There are other strengths: existing buildings can be treated in the same manner as new buildings because detailed information on the building envelope, materials, and heating systems are not required; a national model can be more practical because government-led housing surveys are stratified random samples, or samples that ensure that underrepresented groups are surveyed more frequently, and taking advantage of aggregated data encompasses the entire population. Furthermore, other uncertainties are eliminated, notably self-selection in participants in archetypal research, for example energy-conscious participants overrepresented in energy efficiency demonstration projects. Further discussion on the methodology for household simulation is contained in Chapter 4.

2.4.4 Statistical modelling and parametric data

This section will explain why household simulation models are limited to data that can pass parametric tests for linear and hierarchical regression modelling techniques and why non-parametric and robust regression techniques were not considered. Household simulation is proposed as using social science quantitative methods that take advantage of the available data as approximately parametric as a multilevel linear modelling approach. However, different techniques, such as analysis of variance and multiple regression which are variants of linear modelling that arose from separate disciplines (Cohen, 1968), should be investigated as simpler alternatives.

In social science, most models considered are from the wide family of linear regression models that are built on parametric data. When considering a mix of individual-level and area-level data, linear models are well positioned to compare statistical techniques that consider a variety of ways of treating predictor variables as either interval, continuous, or categorical. When considering the size of a household as a predictor variable, as is currently done for non-heating end-use energy, it is useful to consider different variables that explain the variance in energy use. Floorspace is a continuous variable, but the treatment of numbers of occupants, for example, can be considered either an interval (4 people are considered to be twice the number of 2 people) or a categorical (4 people are considered separately from 2 people) variable to test our assumptions about human behaviour and amount of energy consumed in non-heating end-uses such as electronics, lighting, appliance use, and cooking. It also allows for discretisation of continuous predictors, or the splitting of the data into categorical or binomial variables to test a variety of assumptions.

Non-parametric tests that consider a continuous outcome (e.g. kilowatt-hours) allow for more types of datasets to be used, but there are drawbacks that led to their exclusion from analysis techniques in this thesis. The first and most important of these drawbacks is that these tests use ranked data. This means that the least consuming household would be ranked as 1 with the most consuming household from *n* household ranked as *n.* Any model that is based on ranked data would place all homes between the highest and lowest consuming households at equal distances from each other. This would lose information on any clustering of domestic energy consumption amongst more commonly-found housing sizes (e.g. 3-bed homes are more common than 7-bed homes). Some non-parametric tests are employed in housing stock modelling in the United States (Xiao et al., 2007). More detail on historical changes in the modelling of non-heating end-use energy in the context of the UK is included in Chapter 3.

Further models considered are included in Chapter 4, including statistical techniques including conditional demand analysis, neural network approaches, and engineering techniques such as

distribution of ownership of energy consuming goods in the home. These techniques either had data requirements at the national level that have not been met – for example, the sub-metering of every energy-demanding device to enable the variance of the number of predictors of energy use for different types of non-heating end-uses. This amount of data is available very rarely and when it has been available, the datasets are small and collected in a localised location without overcoming issues of self-selection and non-response biases in the data (Parti and Parti, 1980, Cooke, 2009).

2.4.5 Ecological inference

As aggregate data of both energy use and household size is available through DECC and the UK Census for LLSOAs, researchers might wish to model energy use using average values of small groups. This type of linear regression modelling is called ecological inference, and it is a technique that can be easily misused and misinterpreted (Openshaw, 1984, Freedman, 1999). It has the considerable drawback of modelling the energy-consuming behaviour of an individual household using the mean values of a group of households. However, the practice of data-gathering for building simulation models with the inclusion of innovation parks or energy-efficiency measures in social housing has limited the homes available when testing technological improvements (Dickson et al., 1996) because of the understandable costs of constructing dwellings as part of control groups. This thesis will assess the applicability of ecological inference and the assumptions that need to be made using regression techniques.

Ecological regression has been considered under restrictive criteria to infer individual behaviour from aggregate data when public interest in having analysis is high, where individual data is restricted by privacy considerations, the data at individual level is considered to be low quality, and when a constancy assumption is made (Freedman, 1999, Wakefield, 2004). Ecological regression techniques have been used extensively in political science and in several court cases involving redistricting and testing of voting rights legislation in the United States (Schuessler, 1999).

This thesis contends that these tests can be passed and household modelling can be done as ecological inference. Data protection considerations restrict the ability of the researcher to identify the dwelling to which any housing survey pertains (Great Britain Parliament, 1998). Collection of energy use in households is occasional (Department for Communities and Local Government, 2010a), whilst end-use metering of appliances, electronics, cooking, and lighting is extremely limited in scope and the use of secondary data created as a by-product of studies in the effectiveness of investment in on-site renewable energy generation capacity, raising again the problems of self-selection of householders (Firth et al., 2008, Brown et al., 2007). The scope of work in extensive end-use metering is to test the effect of, for instance, installing appliances with different energy ratings

and the effectiveness of low-energy lighting, and not area characteristics and socio-ecomomic background. Further detail on the appropriateness of the data for this methodology will be provided in Chapter 4.

2.5 Motivations and background summary

This chapter has reviewed the reasons for a new focus on estimating non-heating end-use energy in residential dwellings and identified the need for a methodology for identifying area-level predictors of this type of energy in concert with, or possibly entirely replacing, individual household-level predictors to reduce the gap between predicted and measured non-heating end-use energy. This has been identified as a weakness in current modelling of domestic dwellings, and the models produced by this research should reduce this gap.

First, the model should cover all households, and not limit itself to new build households. The treatment of the model as a model of a household and not a dwelling reduces information requirements such as building materials, sizes of windows, and number of light fixtures that are present in detailed building simulation models.

Second, the model should provide an assessment of the impact of group membership of households, notably area membership. This assessment has a dual purpose of predicting changes in household energy use in large-scale changes to the socio-economic and built environments of neighbourhoods.

Third, the model should be able to respond to the rapidly changing nature of use of appliances, lighting, electronics, and cooking than the current 5- to 10-year cycle of building simulation models. It should make interim changes to its algorithm of non-heating end-use energy demand of households between major housing surveys using aggregate data, whilst the split between end uses and fuel sources is maintained, that is updated annual instead of survey data of energy meters that is updated on an occasional basis.

In addressing the criteria above, a household model of non-heating end-use energy will provide a valuable new source for reducing the gap between predicted and actual energy use, encourage attention to reducing demand in high-electricity consuming households and their neighbourhoods, and more accurately prepare for the transfer of heating end-use energy consumption from natural gas onto the electrical grid. The following chapters will highlight the range of work that has been done in the UK and international contexts in modelling non-heating end-use energy of households as dwellings in building simulation and as part of the residential sector in housing stock modelling, the data available, and the methodologies available for estimating household-level outcomes from large

scale household survey data. These will be drawn upon later to specify a methodology that satisfies the criteria outlined above.

Chapter 3 - Related Work in the context of England

3.1 Introduction

This chapter will summarise the history of single building energy simulation and the assessment of energy consumption using area-level variables in the context of the United Kingdom. The history of building simulation in the UK is dominated by the Building Research Establishment, later renamed as BRE, and their Domestic Energy Model (BREDEM). This model has been adopted by the UK government and its devolved nations for the assessment of potential carbon emissions of every new residential development since 1985 and renamed by the government the Standard Assessment Procedure (BRE, 1998). This section will also explore the practice of validating single dwelling simulation models through housing stock models of the entire residential sector of a single geographic area. Finally, this section will explore earlier work on forming area-level predictors of non-heating end-use energy and the formation of area classification systems, and make recommendations for the direction of research from these findings.

3.2 Domestic energy model

A domestic energy model was defined by the Building Research Establishment in 1985 (Anderson et al., 1985a, BRE, 1998) as a simplified way to calculate the energy use of a dwelling. The model has been defined as a thermodynamic flux, or heat balance, equation built on physics-based research in buildings (Swan and Ugursal, 2009). It has also been defined as a physical-technical-economic model, or a "simple mechanical model used in [energy efficiency] policy and evaluation that focuses narrowly on devices, prices and rationalized behaviour" built on the need for government to measure the impact of policies (Lutzenhiser et al., 2010). The history of domestic energy models in the context of the United Kingdom needs to be viewed in light of these definitions.

The origination of domestic energy models was, according to Shorrock and Anderson (1995) to develop "an energy calculation method...which was substantially better than...design heat loss calculations, but less complex to use than detailed simulation methods". What became known as BREDEM is an energy consumption simulation of a single dwelling created as a paper-based

algorithm due to the limited nature of computing availability in the 1980s. By the end of the 1980s, computer implementations of BREDEM had become the widespread with affordable access to personal computers (Shorrock and Anderson, 1995).

This chapter will explore the development of the modelling of non-heating end-use energy since 1980 in the United Kingdom. The estimation of non-heating end-use energy has been a secondary focus of modelling. The first iterations of the model restricted themselves to estimating heating end-use energy, and even more specifically to a "heating season" defined as being between October and April (Uglow, 1981). The presence of non-heating end-uses were treated as a second order correction to a system dominated by flows of space heating (Anderson et al., 1985b). The estimates of non-heating end-uses were partly based on a count of appliances assumed to be in place in a particular household, and partly based on the size of the household. In later versions, any assumptions about the occupants of a dwelling were removed (Energy Advisory Services, 1996, Jones, 2000), and all predicted non-heating end-use energy are now correlated with the size of the household (BRE, 2010).

Building regulations that promote the conservation of energy use in domestic buildings in the United Kingdom are required under the building acts of its constituent nations as well as fulfilling the European Directive on the Energy Performance of Buildings. (Department for Communities and Local Government, 2007b, Scottish Executive, 2003, Great Britain Parliament, 1984) For simplicity, this thesis will limit itself to the legislation and the current building regulations in place in England and Wales, which are supplemented by approved documents that provide supplementary guidance for fulfilling the requirements of the building regulations. For domestic buildings, Approved Document Part L1A requires a target emissions rate (calculated as the number of kilograms of carbon dioxide per square metre of usable floorspace) that is calculated from the Standard Assessment Procedure (SAP).

3.3 The context: energy crisis, security, and poverty

Energy modelling of dwellings emerged in response to energy crisis, security, and poverty. The energy 'crisis' was built around two spikes in the price of oil in 1973-4 following the Yom Kippur War and the OPEC oil emargo, and in 1979 following the Iranian Revolution. In nominal terms, the rise in energy prices rose threefold from 1970 to 1980 (Department of Energy and Climate Change, 2009a). This raised concerns in the UK over energy security, independence, and self-sufficiency. However, Oil and natural gas reserves in the North Sea were not sufficient to keep supplies up and prices down in the UK. Energy spending as a high proportion of household income (commonly put at 10%) was first

called “fuel poverty” by the Policy Studies Institute for the Commission of the European Communities (Cooper, 1981). These themes became, in turn, driving forces behind the development of models.

The response to the energy crisis was more active government involvement driven by concerns over energy independence and self-sufficiency (Department of Energy, 1978). There was a concerted response to change the mix of fuel for generating electricity in power plants, including the roles of nuclear and coal power. There also was a drive to limit energy demand. This had different manifestations in the living and working environment of the UK – most famously in a three-day week of electricity supply to the commercial sector in early 1974. In domestic homes, the emphasis was, and had been on energy efficiency and the reduction of energy demand for space heating from the 1950s (Building Research Establishment, 1956, Building Research Establishment, 1976, Watson, 1979).

Domestic energy modelling later became a way of measuring warm houses, leading to the concept of the “warm, well-insulated home” (Henwood, 1997, Department of the Environment, 1996). Fuel poverty was defined in 2001 by the Warm Homes and Energy Conservation Act (WECA) 2000 as a person who is a “member of a household living on a lower income in a home which cannot be kept warm at reasonable cost” (Great Britain Parliament, 2000). This was later interpreted by Boardman (1988) as the spending of more than 10 per cent of all household income on heating fuel, leading to modelling that not only estimated on scales based both on the amount and cost of energy, specifically energy for space heating.

These considerations focused the research community on the modelling of the heating end-uses of a dwelling as opposed to the non-heating end-uses. One of the consequences of this was the viewing of the importance of non-heating end-uses as waste heat from the operation of lights, appliances, and mechanical ventilation. The amount of electricity used by these end-uses was not seen as critical to the energy efficiency of UK homes.

3.4 Initial development

3.4.1 BREDEM-1, 1981 and introduction to conditional demand analysis

In 1981 a model emerged in building technology and science publications developed by the mathematician Christine Uglow at the BRE, where thinking on the estimating of domestic heating requirements had been taking place since the 1950s (Building Research Establishment, 1956, Building Research Establishment, 1976, Romig and Leach, 1977). This model later became known as the BREDEM-1 model. BREDEM – the Building Research Establishment Domestic Energy Model - is a

method for estimating the energy used in dwellings for the provision of space and water heating, cooking, lights and appliances. The paper written by Uglow later became known as BREDEM-1. It was a simple single-zone method using seasonal averages for internal and external temperatures (Uglow, 1981).

The first version of BREDEM was created as a heat balance equation, and total non-heating energy end-uses was a predictor variable for heating end-uses and not an outcome variable in and of itself. This was founded on the view that the modelling of the energy use of a residential dwelling, in the words of Uglow, is "a simple equation ... used to estimate the daily energy balance of the dwelling." It was created to assess the change in energy consumption and indoor temperature with energy-saving installations in the building.

Non-heating end-use energy was modelled, even though it was not explicitly called this in the literature, using conditional demand analysis of the number of devices in the home as detailed in Table 3.1. Conditional demand analysis expresses the non-heating end-use energy as a summation of the energy consumed by each of the energy-using devices present in the household that are not part of the heating system. Thus, the non-heating end-use energy consumption of a household is computed as directly related to the appliance stock present in the dwelling. The Electric Power Research Institute in the United States defined conditional demand analysis in the following equation (Electric Power Research Institute, 1989),

$$HEC_{it} = \sum_{j=1}^{j} UEC_{ijt} \times S_{ij} \quad (2)$$

where HEC_{it} is the total non-heating end-uses energy consumption by household *i* in period *t*, UEC_{ijt} is the *j*-unit end-use energy consumption of household *i* in period *t*, and S_{ij} is a binary predictor of household *i*'s ownership of device *j*.

To develop a conditional demand model, household non-heating consumption data HEC_{it} was obtained for a household *i* over survey time or billing cycle *t* in homes that did not use electricity for heating and the overall prevalence of the appliance stock S_{ij} of information obtained through a survey of the household's appliances.

Unit Energy Consumption UEC_{ijt} is a function of the features of household *i*'s energy consuming unit *j* (AF_j) and the utilisation pattern UP_{ijt} that relates to energy-using device *j*. The utilisation pattern of the device itself is a function of the structural aspects of the dwelling ST_i, weather conditions WC_{it}, market conditions MC_{it}, and the socioeconomic situation of the household SEC_i. Thus UEC_{ijt} can be expressed as a function:

$$UEC_{ijt} = F_j(ST_i,\ AF_j,\ WC_{it},\ MC_{it}, SEC_i) \quad (3)$$

Equations adapted from (Aydinalp-Koksal and Ugursal, 2008)

The evaluation of the terms in domestic energy models of non-heating end-use energy using conditional demand analysis usually are done through multiple regression, and not all of the predictors in the unit energy consumption equation are used for every type of non-heating end-use. In the first BREDEM model, the only predictor variable of appliances is the appliance feature predictor – in essence, the rate at which each appliance consumes electricity in a year. Therefore, the unit energy consumption for appliances is given as $UEC_{ijt} = F_j(AF_j, UP_{jt})$ where UP_{jt} is the intensity of use of appliance *j* during year *t.*

The binary S_{ij} in BREDEM is a function of the socioeconomic background of the household that results in a package of 'essential' and 'luxury' goods. This implicitly is results in the function $S_{ij} = f_j(SEC_i)$.

Table 3.1: BREDEM-1 Model Assumptions for annual electricity consumption of appliances and lighting

Electricity consumption, cooking	1190 kWh/annum
Gas consumption, cooking	2380 kWh/ annum
Electric appliances (basic basket of goods)	912 kWh/ annum
Electric appliances (luxury basket of goods)	2555 kWh/ annum
Lighting	292 kWh/ annum

(Uglow, 1981)

In a follow-up paper, Uglow claimed that this approach proved capable of producing realistic estimates of heating end-use energy in real residential dwellings. However, when studying the non-heating end-uses by measuring electricity in homes that did use electricity for heating end-uses, Uglow reported that predicted values of consumption based on conditional demand analysis of electrical supply statistics were found to be low. Unfortunately, in this paper the full analysis of non-heating end-use energy was not reported (Uglow, 1982).

3.4.2 Second generation of BREDEM: BREDEM-2 to BREDEM-7

The second generation of BREDEM models (1983-1986) moved to a two zone model in order to describe the heating of main living and other spaces to different temperatures. BREDEM-2 allowed for wider range of heating system types in addition to natural gas boilers and introduced the variable degree-day approach to external temperatures for evaluating energy demand on a yearly basis instead of a standard heating "season". Further variants are as follows:

- BREDEM-3 for the 'Energy Matters' home energy audit scheme (Alexander, 1983). The report by Anderson et al (1985) pulled together the requirements for BREDEM-2 and 3. BREDEM-3 was the first edition of BREDEM to be designed with algorithms instead of look-

up tables for analysis using a computer. It was designed for the 'Energy Matters' television show produced by the Open University for Channel 4 that prompted viewers at home to fill out questionnaires about their homes submitted for analysis at the BRE. Advice on the types of energy-saving installations that would reduce heating end-uses was then given to households that filled out the questionnaire.

- BREDEM-4 for the Home Energy Audit Advice and Treatment (HEAT) scheme where surveyors were to collect information on homes that were for sale. This model was never implemented.
- BREDEM-5. The energy designer model (Chapman, 1990). This is a commercial version of the BREDEM model used for the Milton Keynes Energy Cost Index (MKECI). It was the forerunner of labelling schemes and used for the Energy World Exhibition of 1985 and applied to all Energy Park developments in MK. Monitoring proved its predictions to be reliable. (Shorrock and Anderson, 1995)
- BREDEM-6. The energy auditor model. The predictive capabilities of the two-zone version (i.e. BREDEM-2) were tested and some improvements were identified for use in later models and submitted for publication in the Journal of Building Services Engineering Research and Technology (Henderson and Shorrock, 1986b).
- BREDEM-7. Energy Efficiency Office running cost guide – the Monergy Guides in mid 1980s.

The algorithms in BREDEM-3 are the most widely circulated because of the model's connection with the Open University and wide dissemination and publications, notably a full document on the background, philosophy, and description of the BREDEM model released by the BRE in 1985 (Anderson et al., 1985b). It produced Tables 3.2 to 3.4 below of appliances and lighting to be modelled using conditional demand analysis. All tables are adapted from Anderson et al. (1985b)

The unit energy consumption of lighting is a function of the structural characteristics of each room, the room features (the type of room and its floor area), and the socioeconomic situation of the household (if there are children present). It also introduces a concept of diminishing returns of energy use as the size of the household increases as detailed in Table 3.2.

$$UECL_{ijt} = F_j(ST_i,\ AF_j,\ SEC_i) \quad (4)$$

Table 3.2: BREDEM-3 estimates of appliances and lighting demand of dwellings by room

Zone	If floor area >= 200 sqm	If floor area < 200 sqm	If children=yes, then add additional
Living room	175.0 kwh/annum	0.9 (floor area) kwh/a	52.5 kwh/annum
Bedrooms	43.7 kwh/annum	0.2 (floor area) kwh/a	13.1 kwh/annum
Rest of house	131.2 kwh/annum	0.7 (floor area) kwh/a	39.4 kwh/annum

The unit energy consumption of cooking is given as a constant, depending on the source of fuel:

Table 3.3: BREDEM-3 estimates of cooking demand by fuel type

Fuel	Consumption
Electric	944.7 kwh/annum
Gas	1189.6 kwh/annum

The unit energy consumption of appliances is assumed to be a function of the appliance features (the rate of consumption per hour and the number of hours of use per day). The number of appliances in the household is taken to be a given in the model, with each assigned an electricity consumption value in Table 3.4. At this time, BREDEM was a model built to predict the energy use of existing homes, represented by the algorithm

$$UECA_{ijt} = F_j(\, AF_j) \tag{5}$$

Table 3.4: BREDEM-3 estimates of appliance demand by device

Appliance	Electricity consumption
Refrigerator	262.4 kwh/annum
Kettle	174.9 kwh/annum
Freezer	673.5 kwh/annum
Television	236.1 kwh/annum
Washing machine – Hot fill	43.7 kwh/annum
Washing machine – Cold fill	157.4 kwh/annum
Dishwasher	271.1 kwh/annum
Tumble Drier	148.7 kwh/annum
Miscellaneous	122.4 kwh/annum

3.4.3 Validation, overestimation claims, and the emergence of the third generation of BREDEM

During the rest of the 1980s, the Building Research Establishment set about collecting data that would validate this model of energy use, with the key metric being the variance between the totals for the actual versus predicted amounts of consumption combining both heating and non-heating end-uses. The data obtained from field trials (Building Research Energy Conservation Support Unit, 2007) by Henderson and Shorrock did not agree with the BREDEM-3 estimates. Instead, Henderson and Shorrock determined that the product of the number of occupants and the total floor area was the best predictor of non-heating end-use energy (Henderson and Shorrock, 1986a). The paper proposed a relationship between the product of total floor area and the numbers of occupants as follows using a second-order polynomial. This became part of later versions of BREDEM:

$$ELEC = 2693 + 4\,(TFA \times N) - \,8 \times 10^{-4}(TFA \times N)^2 \tag{6}$$

where *ELEC* is annual electricity consumption in kilowatt-hours, *TFA* is the total floor area in square metres, and *N* is the total number of occupants, which can either be known or derived from total floor area.

However, in the creation of the Milton Keynes Energy Cost Index (MKECI), Chapman claimed that this method led to an over-estimation of energy use in larger homes. The Milton Keynes Energy Cost Index was an a measurement standard used by the Milton Keynes Development Corporation to ensure dwellings have a better energy performance than required by building regulations in place at the time. The standard was applied to 3,000 dwellings after initial development for the "Energy World" exhibition of fifty homes in 1985. The standard was designed to run on a standard personal computer. Instead, Chapman proposed a straight linear relationship between electricity use and usable floor area (Chapman, 1990):

$$ELEC = 22.22 \times TFA \quad (7)$$

where *ELEC* is annual electricity consumption in kilowatt-hours and *TFA* is the total floor area in square metres.

However, both researcher groups acknowledged that there were other moderating factors. Both of these models allowed for a significant reduction in electricity demand reduction from low-energy (fluorescent) lighting using a new variety of the conditional demand analysis (CDA) model. Henderson and Shorrock stated that the reduction could be as much as 80 per cent if all rooms employed such devices; Chapman assumed only a 50 per cent reduction, but this figure was based on the presence of any low-energy lighting in the builder's construction specification. In addition, the use of household income as a predictor of energy use was discussed, but not brought forward in the algorithms developed by either research group.

Both of these approaches deviated from the initial domestic energy models for non-heating end-use energy because they did away with the conditional demand analysis (CDA) approach that involved the estimation of the number of appliances, and instead introduced two different varieties of linear regression models. Therefore, the emerging formula for non-heating end-uses became one that included assumptions on the performance of lighting, with the assumptions on the occupation of the home transformed from a binary variable of the presence of children to one derived from the physical size of the household $\{N_i = f_i(TFA_i)\}$:

$$UEC_{ijt} = F_j(ST_i,\ AF_j,\ UP_{jt}) \quad (8)$$

This intermediate period signalled a new evolution in the modelling of non-heating end-use energy as BREDEM became used as part of the checking of building regulations. In order to do so, many of the detail of the previous versions of BREDEM, notably the list of appliances owned and operated by the occupants, were dropped in favour of an approach that positioned domestic energy modelling as

part of the building regulation procedure. This was part of a wider trend for performance-based as opposed to prescriptive regulation (Inter-jurisdictional Regulatory Collaboration Committee, 2010, Meacham et al., 2005). In building regulation, the assessor only needed information from the housing provider about the floor area of the building and the types of lighting fittings that would be installed at the start of occupation with the assumption that they were unlikely to be removed in favour of a different lighting fixture type. This would lead to a move from research in the effect of new installations on the heat balance of the building, to regulating the designed energy performance of the building, after the Building Act 1984 permitted the Secretary of State to enact building regulations in England that encompassed "the conservation of fuel and power." (Great Britain Parliament, 1984)

3.5 Regulatory focus and the third generation of BREDEM as the first generation of the Standard Assessment Procedure

The Standard Assessment Procedure was first published by the then Department of the Environment with the Building Research Establishment in 1993 and in amended form in 1994, and subsequently the algorithms were published as conventions in 1996 and amended in 1997. Revised versions were published in 1998, 2001 and 2005. SAP was integrated into the building regulations of England and Wales through Regulation 14A of the Building Regulations (Amendment) Regulation 1994 and published as guidance in *Approved Document Part L : Conservation of Fuel and Power* (referred to in architectural and engineering practice as Part L) in 1995.

After the passing of the Building Act 1984, the first edition of Part L was published in 1985, replacing guidance in Part F of the building regulations. This detailed only one compliance approach which addressed space heating only, which it called the Elemental Method, which specified maximum thermal transmittance levels for each building's external and internal walls and windows in terms of a U-value.[1] The 1989 revision to the building regulations provided for an Energy Target method of compliance with a predicted energy demand less than the elemental method, as "this method can allow for useful heat gains" (DoE, 1990). The 1995 revisions to Part L abolished the Energy Target method and replaced it an energy rating method. The energy rating method for dwellings was the Standard Assessment Procedure (SAP). A further development of this rating method became mandatory for all dwellings with the revision of Part L in 2006 (Department for Communities and Local Government, 2006a).

[1] U-values indicate thermal transmittance: how much heat will pass through one square metre of a structure when the air temperatures on either side of a structure differ by one degree Kelvin. They are expressed in units of Watts per square metre per degree of temperature difference $Wm^{-2}K^{-1}$ Defined in CIBSE 2006. *Environmental design : CIBSE guide A*, London, Chartered Institution of Building Services Engineers.

3.5.1 Principles of the Standard Assessment Procedure (1995-2010)

The Standard Assessment Procedure (SAP) was created in the mid-1990s and changed the nature of the process from estimating energy use in existing buildings to estimating the potential energy use of new buildings in order to grant building permission. SAP in 1995 retained the previous focus on heating – the final rating is a ratio of heating costs over household size.

SAP changed the criteria on which non-heating end-use energy could be estimated. No longer was it allowed for the assessment criteria to make any assumptions about the potential or current occupants of a dwelling, nor the location of the dwelling (Energy Advisory Services, 1996, Jones, 2000, Department for Communities and Local Government, 2006a). This meant that any previous assumptions about the ownership of appliances in a basic or luxury basket as first proposed in the first generation of BREDEM could not be used. Fundamental to the adoption of SAP was the assumption that people could be reduced to a physical element of the dwelling (Energy Advisory Services, 1996, Jones, 2000). The removal of occupant behaviour effects was a requirement of the government for the adoption of SAP, and this approach fits into an overall physical-technical-economic model of energy use(Lutzenhiser et al., 2010). The predicted electricity use for non-heating end-use energy was related to the size of the household, treating the house as a singular appliance *j* subject to the size of the household *i*:

$$UEC_{ij} = F_j(ST_i) \tag{9}$$

where the structure ST_i of the dwelling is represented by the total floor area of the building. The basic SAP algorithm in place from 1995 to 2010 for appliances, cooking, metabolic gains, and lighting, was as follows (BRE, 2001):

$$N = (0.035 - TFA) - (3.8 \times 10^{-5} \times TFA^2) \text{ if } TFA \leq 420 \tag{10a}$$

$$N = 8.0 \text{ if } TFA > 420 \tag{10b}$$

$$W = 74 + (2.66 \times TFA) + (75.5 \times N) \text{ if } TFA \leq 282 \tag{11a}$$

$$W = 824 + (75.5 \times N) \text{ if } TFA > 282 \tag{11b}$$

Where *N* is the number of occupants, *TFA* is the total floor space in square metres, and *W* is the annual amount of free heat gained in the indoor environment from internal sources in watts.

3.5.2 BREDEM-8 and BREDEM-12 – the "baseline"

In 1996, the third generation of BREDEM crystallised into the BREDEM-8 and BREDEM-12 models. BREDEM-8 was a domestic energy model that was designed to calculate the energy demand for a

given month of the year taking account the effects of the strength of the sun on solar gain and the outdoor air temperature influencing heating demands in that month (Anderson, 2002a). BREDEM-12 is a calculation of the annual amount of energy needed to service a dwelling, using the concept of the heating season (now defined as October to May) (Anderson, 2002b). This section focuses on the "baseline" calculation: this assumes no knowledge or assumptions about the behaviour of occupants of the dwelling in relation to lights, appliance ownership, or cooking.

BREDEM-8 and BREDEM-12 were the first models designed for calculation on the computer and the available complexities that could go along with this. They paralleled an International Standard, ISO 9164, for the calculation of energy use in dwellings (International Organization for Standardisation, 1989) – although this deals with space heating only. Work subsequently began on a European Standard which considered some additional aspects of energy calculation – although this paper again specifically refers to the energy requirements for heating only (European Committee for Standardization, 1992).

The BREDEM-12 model is the preferred model to examine in this thesis because it predicts annual energy use. The two different interpretations of Henderson and Shurrock and Chapman in BREDEM-5 and BREDEM-6 are retained, and it is left to the assessor which approach they wish to use. The first formula follows Chapman, and instructs the assessor to use it "in order to avoid the rating being dominated by electricity consumption", supposedly in response to Chapman's concerns of over-estimation $\{E_{LA} = 24 \times TFA\}$. The Henderson and Shorrock equations are provided "if a more accurate prediction of electricity use is required:"

$$E_{LA} = 619 + 6.44\ (TFA \times N) \text{ if } TFA\ \times N < 710 \quad (12a)$$

$$E_{LA} = 2700 + 4.05\ (TFA \times N) - 0.214\ (TFA \times N)^2 \text{ if } 710 \leq TFA\ \times N < 2400 \quad (12b)$$

$$E_{LA} = 7990 \text{ if } TFA\ \times N \geq 2400 \quad (12c)$$

where E_{LA} is the electricity consumption for appliances and lighting in kilowatt-hours per annum, *TFA* is the total floor area in square metres, and *N* is the calculated number of occupants dependent on floor area (Anderson, 2002b)

Cooking end-use energy is calculated based on the fuel for the cooking appliance installed (all-electric, all-gas, hybrid gas hob and electric oven) and the numbers of people. In the case of the all-electric cooking system, $\{E_k = 472 + 94.4 \times N\}$ where E_k is the electric cooking consumption in in kilowatt-hours per annum.

These algorithms are part of the general movement of the BREDEM method from the mid-1980s from a conditional demand analysis based on ownership of appliances to a linear regression model based on household size. However, it emerges from the research on validation that the models were driven by the available data now known as the monitored energy use archive, which did not take an inventory of appliances or the socioeconomic circumstances of the occupants, since the focus of the investigations was on the reduction of energy use through better space and water heating energy efficiency installations (Building Research Energy Conservation Support Unit, 2007). People became represented in the BREDEM model as a physical element of the dwelling as there was no information about the people who might inhabit it.

3.5.3 Underestimation, overestimation, and verification of the third generation of BREDEM

There was a series of articles published between 1991 and 1996 that detailed the testing the model against the data available from energy efficiency projects, notably in Wales, Birmingham, and Washington, Tyne and Wear. The third-generation BREDEM models were tested against real measurements of dwellings and against other detailed simulation models as presented in Building Environmental Performance Analysis Club conferences in 1991 and 1994 at Canterbury and Kent and some subsequent journal articles (Shorrock, 2010). There were also papers published internally as additional supporting evidence to BRE. These papers constitute the publicly released evidence for the third generation of BREDEM and the first generation of SAP.

The first working paper on data available to support the BREDEM/SAP model in use from 1996 to 2010 was presented at the 1991 Building Environmental Performance Analysis Club Conference in Canterbury (Shorrock et al., 1991). The amount of data taken to support the model was 222 dwelling years from 155 dwellings in seven different data sets where some measurement of the main-end uses of space heating, water heating, cooking, and lights/appliances were available. The target amount of data was the equivalent of 200 dwelling years. After use in verifying the model, Shorrock et al. found that not all of the datasets were ideal for the purpose of BREDEM.

Table 3.5: BEPAC assessment of suitability of available datasets in the formation of BREDEM-8/12

Dataset	Description	Suitability as assessed by the investigators
BRE	Standard BRE dataset of 4 passive solar homes at Linford and 10 other homes	Ideal
Washington, Co Durham	6 Low energy, well insulated terraced homes	Not suited
Birmingham Energy Improvement Kit	Birmingham City Council housing project to install a new kit in 25 older dwellings with large amounts of social data	Not ideal
Sandwell, Birmingham	Energy efficiency demonstration project in a high rise block heated by electric storage	Not ideal
Collyhurst, Manchester	Demonstration project with 30 homes with gas fired condensing boilers	Ideal
Milton Keynes Energy Park	25 homes. Designated in 1985, the Energy Park was planned as an international demonstration project of energy efficiency. All buildings constructed in the Energy Park were required to demonstrate high levels of energy efficiency.	Ideal
Super Insulated, Milton Keynes	4 super-insulated and 4 other homes at Two Mile Ash in Milton Keynes. This was a demonstration project funded by the European Commission and the Polytechnic of Central London (now Westminster University)	Not ideal

The investigators made some commentary on the problems in estimating energy use for non-heating end-uses. They asserted that the most effective way of measuring this amount of energy use was to measure behaviour, using the appliances present as part of a conditional demand analysis model . The problems that BRE encountered at that point in testing was when the product of floor area and occupants was large, the model over-predicted, and when the product was small, the model under-predicted, specifically in the case of a single occupant of a one-bedroom flat (Shorrock et al., 1991).

The builders of the BREDEM model returned to the next Building Environmental Performance Conference in 1994 with more comparisons between measured and predicted energy use in dwellings using BREDEM/SAP (Shorrock et al., 1994). In this review, the team used only “data of the highest quality” from the Milton Keynes Energy Park project. There were 19 dwellings with central

gas heating and gas cookers selected with 2 years of monthly data each. In these dwellings, an excellent agreement was found between overall predicted and measured energy use. In the monthly model, BREDEM-8, the gradient was 0.94 and the correlation coefficient was 0.95. In the annual model, BREDEM-12, the gradient was 0.96 and the correlation coefficient was 0.95. No breakdown of heating and non-heating end-use energy was presented in the paper.

The "baseline" calculation of non-heating end-use energy was deemed to be insufficient when matched up against the actual measures of energy use for the purposes of pure research, but sufficient for the purposes of building regulation. This led to two algorithms: one BREDEM algorithm for use in the buildings research community; and a second for calculating energy ratings for the checking of building regulations.

3.5.4 Retaining of conditional demand analysis in the third generation of BREDEM

A supporting evidence paper (Energy Advisory Services, 1996) was written to detail the evidence and discussion around the third generation of BREDEM during the 1990s. It details the pressures that the BREDEM team came under from academic research to include socioeconomic indicators such as income in the prediction of non-heating end-use energy. It also details the large, and seemingly, random variance of this type of energy consumption that cannot be explained physically. Finally, the document details the final decisions made by the BRE to allow energy assessors to assign a multiplier or divisor to the modelled consumption based on socioeconomic assumptions, with the basis for those assumptions based on *a priori* knowledge.

The results shown in the Pennyland Project, an experiment on an estate of 177 houses in the Pennyland district of Milton Keynes to measure the impact of passive solar design, indicate that the 'gap' between measured and estimated electricity use was most severe at for the smallest-consuming households. (Chapman et al., 1985b) Another monitoring project at Linford, UK also posited that the best explanation of the variation in non-heating end-use energy was household income (Everett et al., 1985).

In addition to the specific problems associated with low income households, energy monitoring projects found a large spread of 'average' households in non-heating end-use energy as represented by electricity use in homes with natural gas heating fuels. The seven houses at Linford with identical houses with identical numbers of occupants had electricity consumption in appliances ranging from 1805 kWh/yr to 4289 kWh/yr. The range from lowest to highest use of electricity used for cooking was from 446 kWh/yr to 1097 kWh/yr. The standard error implied by these results was about 25% of the mean. (Everett et al., 1985) The Pennyland Project was quoted by a supporting paper for

BREDEM (Energy Advisory Services, 1996) to have a standard error around 10% - but it may not be a correct interpretation of the data presented.

When faced with the task of making a proposal to alter the model to take into account income, Energy Advisory Services (1996) explained that "it is assumed that it is unacceptable to ask households for an estimate of their disposable income." Nevertheless, surveyors were expected to be able to determine if households were in higher or lower income bands. The assessment for input into the BREDEM model is based on looking at the overall level of appliance ownership and "indications of the level of household income."

The surveyors were then instructed that around 70 per cent of all homes fit into the 'average' band with no alteration of the BREDEM algorithm, and that they should be placing about 15% households into cases higher or lower than average. Households could be 'well below average' in exceptional circumstances, and only for those in fairly extreme poverty, typically with no income other than benefit payments. The algorithm could then be recommended to be re-scaled as follows for electricity use in lights, appliances and cooking:

Table 3.6: Recommended re-scaling of appliances and lighting in BREDEM-8/12

Average case	**as algorithm**
Higher than average	+ 20%
Lower than average	- 20%
Very low use	- 40%

(Anderson, 2002a)

Therefore the conditional demand analysis model still remained in a reduced format in the third generation of BREDEM. The appliance features of low-energy lighting were taken into account, and the amount of lighting consumption could be reduced if there is documented evidence of low-energy light fixtures installed. In addition, a socioeconomic factor could also be assumed to affect the usage pattern of lights, appliances, and cooking. Therefore, the following estimate of the unit energy consumption for non-heating end-use energy returned to one that contained elements of conditional demand analysis:

$$UEC_{ijt} = F_j(ST_i,\ AF_{jt},\ UP_{jt},\ SEC_i) \qquad (13)$$

Where ST_i is the structural aspects of the household, represented by household size in total floor area, AF_{jt} are the appliance features of low-energy light fixtures, represented by the number of zones that had these light fittings, UP_{jt} is the intensity of use of appliance *j* during year *t,* and SEC_i are the socioeconomic charactertistics of the household, which are represented by the household

tenure, either known or assumed for the area in which the household was located. The next section will explore the types of assumptions that could justify lowering or raising the scale of non-heating end-use energy according to the BREDEM model used from 1996 to 2005.

3.5.5 The emergence of bottom-up housing stock conditional demand analysis models: DECADE

During the 1990s, a new, bottom-up housing stock modelling of non-heating end-use energy, particularly the electricity use of lights and appliances began to emerge. Bottom-up modelling, was, according to (Swan and Urgusal, 2009), "developed to identify the contribution of each end-use towards the aggregate energy consumption value of the residential stock." Modellers separated the heating and non-heating parts of the energy use of the domestic sector (Guler et al., 2001). In one example (Bennett and Newborough, 2001) the energy use attributed to dwellings was entirely due to its "heating efficiencies" and length of the heating season in days. The energy use of appliances and lighting in the residential sector had nothing to do with the nature of households, but was determined by annual energy consumption by appliance type and usage pattern, described previously as the unit energy consumption in a conditional demand analysis model.

A new strand of conditional demand analysis modelling of the entire residential stock emerged in the early 1990s in a major study on wet appliances commissioned by the Directorate General for Energy of the European Commission and led by the Danish Energy Agency (Group for Efficient Appliances, 1995). Based on the work of a team at the Oxford University Environmental Change Institute (Hinnells et al., 1995) a consensus was built around so-called "vintage" models of the energy consumption demand of lights and appliances (Jacobsen, 1998). This included a detailed stock turnover model to simulate the scrapage, replacing, and addition of appliances with different energy efficiencies. Figure 3.1 shows an example a report written by Sustainable ENvironment COnsultants (SENCO) of the change in dishwasher stocks as the ownership level rises and reaches market saturation and then slows to the same rate of increase as household formation.

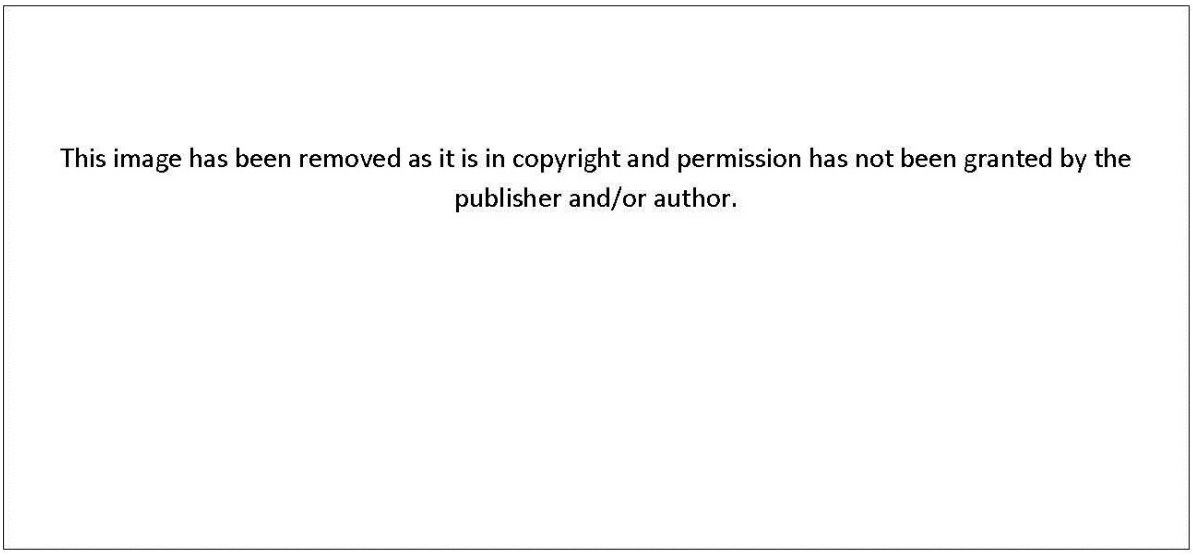


Figure 3.1: Domestic dishwasher stock model in SENCO (2005)

These vintage models depended on the detailed knowledge of a large number of end-use devices in the market relevant to the housing stock model. The penetration rate for each technology, such as an electric appliance, is described with a saturation level with a different consumption level for the year of vintage in the year being investigated. Therefore, the unit energy consumption for all appliances in the housing stock in a year is:

$$UEC_{ijt} = F_j(AF_{ij},\ UP_{jt}) \quad (14)$$

Where AF_{ij} represents the electric consumption of any appliance *j* of vintage *i* (replacing household *i* at the second level of investigation), and UP_{jt} is the intensity of use of appliance *j* during year *t.*

The CDA algorithm for household energy consumption now represents the consumption of all the appliance stock:

$$HEC_{it} = \sum_{j=1}^{j} UEC_{ijt} \times S_{ij} \quad (15)$$

Where the binomial variable S_{ij} is now determined by if appliance *j* of vintage *i* is present in the year of investigation t. This is determined by the algorithm

$$S_{ij} = SALES_{ij}(1 - a_{ij})^{t-i} \quad (16)$$

Where $1/a_{ij}$ is the average lifespan for appliance *j* of vintage *i* and $SALES_{ij}$ is the size of sales in vintage year *i* of appliance *j.*

The DECADE model (Environmental Change Institute, 1995) was a nationwide vintage model built specifically for the United Kingdom that took in data from the sales of lights, appliances, and

cooking. Trends of ownership of appliances and electronics at the UK level were collected by various national agencies, universities, quangos, and charities concerned with energy use. The DECADE project obtained data for the numbers of appliances of different types bought and disposed of per year, assigned a power consumption and use frequency measure, and totalled all of the electricity use together for appliances and lighting and the combined electric and gas use for cooking. Simplified versions of the model, for example without intensity of use of every vintage of every appliance, were also available.

3.5.6 The 'physically derived bottom-up stock model BREHOMES' and rescaling factors – socioeconomic or physical?

The *Domestic Energy Fact File*, produced by the Building Research Establishment, used this as a bottom-up model to estimate the UK-wide energy use of lights and appliances (Shorrock and Utley, 2008). The claim of the *Energy Fact File* was that a bottom-up model built from BREDEM matched the bottom-up model built from DECADE through the stock model called BREHOMES developed in the mid-1990s by the Building Research Establishment(Shorrock and Dunster, 1997, Ekins and Dresner, 2006). BREHOMES created a range of architypes, or typical homes, and used census data on the average sizes of households to construct a breakdown of the numbers of homes nationally that fit into each archetype. The sum of all of the homes' energy consumption for lights and appliances according to the BREDEM model was compared to the total energy consumption of lights and appliances according to the DECADE model for that year. The BREHOMES model assumes the DECADE model to be more accurate, and Table 3.7 details scaling factors introduced in the 1996 version of BREDEM-12 are applied to all non-heating end-use energy (inclusive of lights, appliances, electronics, and cooking):

Table 3.7: Recommended re-scaling in BREHOMES

User	BREDEM Scaling Factor	Starting application assumption
Average case	as baseline	Middle 70 % of dwellings
Higher than average	+ 20% of baseline	Top 15% of dwellings
Lower than average	- 20% of baseline	Bottom 15% of dwellings
Very low use	- 40% of baseline	Marginal

The percentage of the households to which each scaling factor applies were assumed to be around 15% of what are considered to be the richest or largest homes in the archetypal language built into the BREHOMES model. The assignment of homes into "average cases" or cases that were higher or lower than the average was difficult to justify with evidence and the 15 percent scaling factor is

notably larger than the 5 to 10 percent scaling factor found when studying the averages of different area classifications later in this thesis. The evidence for assigning homes to be above or below average is connected with household income, as the criteria for assigning a household to having very low usage was based on income (Energy Advisory Services, 1996). Without this information for new-build homes, both the appliance features and socioeconomic features continued to be excluded from the conditional demand analysis function in SAP as applied to building regulations.

Estimating socioeconomic variables, notably income, in new-build homes to estimate lighting, appliances, electronics, and cooking was not taken any further by researchers because of all the difficulties of convincing government regulators that such an assumption could be safe if challenged (Jones, 2000). The basis for any legal challenge of the algorithm can be found in the developers' own conference papers (Shorrock et al., 1994, Shorrock et al., 1991). The validation of the third-generation BREDEM models did not have data that met parametric tests, and therefore the authors admitted that the values of any parameters estimated from the data were not statistically significant, but that they should be taken forward due to the absence of any other data on domestic energy use. Furthermore, the government was required to developing building regulations that set standards for the conservation of fuel and power, and in the setting of energy targets and ratings it was necessary to move beyond the Elemental Model (Shorrock et al., 1991). Assumptions based on physical parameters were accepted, but assumptions of lifestyle based on tenure or the area the occupants resided in were not acceptable, and explicitly removed from the first editions of SAP (BRE, 1998).

Variation of household energy use based on the housing development or area has yet to be fully explored. With the advent of marketing databases such as ACORN and Mosaic using large amounts of proprietary information, some organisations began to explore these assumptions, the best known being the Energy Saving Trust. The Energy Saving Trust commissioned a market segmentation approach that classified postcodes with data input from the Mosaic classification database and knowledge of the number of heating end-use installations (BRE, 1998, Energy Saving Trust, 2009a). However, the evidence from research performed on behalf of the EST pointed towards a focus on reducing heating end-uses in the targeting of consumer messages in order to change attitudes and behaviours (Darnton, 2006, Moore, 2008) in EST's recent pamphlets about the rise of the use of electronics (Energy Saving Trust, 2011, Energy Saving Trust, 2007a).

3.5.7 Scenario testing of the application of rescaling factors in BREDEM since 1996

Since the first DECADE models used in conjunction with BREDEM and BREHOMES in 1996, the total energy consumption of lights and appliances as estimated by the vintage model DECADE has

increased. The baseline BREDEM-12 model and the SAP models (revised up to 2005) that estimated non-heating end-use energy remained the same. Therefore, the percentage of households that were rescaled up had to increase to match the bottom-up model built from BREDEM to the more 'reliable' model of DECADE, and most probably the percentage of homes rescaled down had to decrease. The exact archetypes used by the Building Research Establishment are unknown, but the BRE's *Domestic Energy Fact File* for 2008 states that non-heating end-use energy increased from 293.2 petajoules per year (PJ/year) in 1996 to 321.2 PJ/year in 2006 in England, an increase of 9.5% overall or roughly 1% per annum (Shorrock and Utley, 2008).

The treatment of this increase in non-heating end-use energy had consequences for the assumptions made in the original 1996 BREDEM scaling algorithm. There were several choices available that can be described without needing to know the exact breakdown of archetypes which have not been published by the BRE – to produce scaling factors that match the 2006 consumption of non-heating end-use energy in England. Below are three plausible scenarios in 2006 deriving from a baseline scenario created for England for 1996:

- **Scenario 1.** Increase the above-baseline consumption of the top 15% of households from 20% above baseline.
- **Scenario 2.** Increase the percent of top-consuming households from 15% of households whilst maintaining their consumption at 20% above baseline.
- **Scenario 3.** Eliminate all bottom-consuming households and redistribute all households to either baseline and 20% above baseline households.

These scenarios are computed from the algorithm

$$E_{total,t} = E_{baseline} H_{total,t} \left(HP_{bottom} + HP_{baseline} + HP_{top}\right)\left(UP_{bottom} + UP_{baseline} + UP_{top}\right) \quad (17)$$

Where $E_{total,t}$is the total consumption of non-heating end-use energy in petajoules in England in year t,

$E_{baseline}$ is the annual mean consumption in petajoules for a million households at the time of the release of BREDEM-12 in 1996,

$H_{total,t}$ is the total number of households in year t in millions,

$HP_{bottom}, HP_{baseline},$ and HP_{top} are the percentages of households below, at, and above baseline, respectively,

and UP_{bottom}, $UP_{baseline}$, and UP_{top} are the percentages of baseline non-heating end-use energy households use below, at, and above baseline, respectively.

Baseline Scenario, 1996

This scenario solves for $E_{baseline}$ with the following values assumed for 1996:

Table 3.8: Baseline scenario testing – starting values

$E_{total,1996}$	293.2 petajoules (PJ)
$H_{total,1996}$	20.2 million households (million hh)
$\boldsymbol{E_{baseline}}$	**14.5 PJ/million hh**
HP_{bottom}	.15
HP_{middle}	.70
HP_{top}	.15
UP_{bottom}	.80
UP_{middle}	1.0
UP_{top}	1.2

The baseline energy use for the residential housing stock of England in 1996 is 14.5 PJ/million households, or 4030 kilowatt-hours per household.

Scenario 1: The top-consuming households use more than 20% above baseline, 2006

This scenario solves for UP_{top} assuming the following values for 2006 and the BREDEM baseline calculated above:

Table 3.9: Baseline scenario testing – solving for top-consuming households use more than 20% above baseline

$E_{total,1996}$	321.2 petajoules (PJ)
$H_{total,1996}$	21.5 million households (million hh)
$E_{baseline}$	14.5 PJ/million hh
HP_{bottom}	.15
HP_{middle}	.70
HP_{top}	.15
UP_{bottom}	.80
UP_{middle}	1.0
$\boldsymbol{UP_{top}}$	**1.4**

Therefore, if the number of homes that are defined as the top-consuming homes remains constant, the amount of energy per household above baseline per household would rise from 20% above baseline to 40% above baseline.

Scenario 2: There are more top-consuming households than 15% of the population, 2006

This scenario solves for HP_{middle} and HP_{top} assuming $HP_{bottom} + HP_{middle} + HP_{top} = 1$ the following values for 2006 and the BREDEM baseline calculated above:

Table 3.10: Baseline scenario testing – solving for top-consuming households comprising more than 15% of the population

$E_{total,2006}$	321.2 petajoules (PJ)
$H_{total,2006}$	21.5 million households (million hh)
$E_{baseline}$	14.5 PJ/million hh
HP_{bottom}	.15
$\boldsymbol{HP_{middle}}$	**.55**
$\boldsymbol{HP_{top}}$	**.30**
UP_{bottom}	.80
UP_{middle}	1.0
UP_{top}	1.2

Therefore the percentage of households that are defined as top-consuming would rise from 15% to 30% of the population and the percentage of average households would drop to 55% if the amount that they consume above average remains constant at 20% and the percentage of low-consuming households remains constant at 15% of the population.

Scenario 3: There are no low-consuming households left in 2006

This scenario solves for HP_{middle} and HP_{top} assuming $HP_{bottom} + HP_{middle} + HP_{top} = 1$ and the following values for 2006 and the BREDEM baseline calculated above:

Table 3.11: Baseline scenario testing – solving for no low-consuming households

$E_{total,2006}$	321.2 petajoules (PJ)
$H_{total,2006}$	21.5 million households (million hh)
$E_{baseline}$	14.5 PJ/million hh
HP_{bottom}	.00
$\boldsymbol{HP_{middle}}$	**.85**
$\boldsymbol{HP_{top}}$	**.15**
UP_{bottom}	.80
UP_{middle}	1.0
UP_{top}	1.2

Therefore, the percentage of high-consuming households would remain constant at 15% and all of the low-consuming households would use non-heating energy at average levels, raising this proportion of the population to 85% if low-consuming households were eliminated and high-consuming households continued to use 20% above average energy use.

3.5.8 Conclusions from scenario testing

From the scenario testing, the scaling factors that BREHOMES applied to non-heating end-use energy deriving from the summing of single-dwelling BREDEM models to match the DECADE model estimate changed considerably from the model's inception in the decade between 1996 and 2006. The underlying algorithm for energy consumption per typical household in BREDEM did not change during this time. A number of different assumptions about the housing stock were likely needed in order to accommodate this lack of change. This work outlined three different ways that the BREHOMES housing stock model could have approached this work through changes to its rescaling procedures.

It also highlights the importance of the definition of the scale of investigation in domestic energy modelling, as these rescaling procedures were not part of the Standard Assessment Procedure for individual buildings. Without rescaling of individual households in a housing stock model, the official estimation of carbon emissions due to lights and appliances in households would have been considerably lower than reality. Still, patterns of consumption by individual dwellings and housing developments are modelled without making any of the socioeconomic assumptions used in the rescaling process in stock modelling, and therefore SAP estimates of energy use of the building design became more and more likely to be lower than the measured energy use once the households were occupied. This likelihood was flagged up by an official review of UK fuel poverty in 2005, and further reviews in the preparation of SAP2009 and DEMSCOT, a new Scottish domestic energy stock model (Sefton and Chesshire, 2005, Henderson, 2009, Scottish Government Social Research, 2009).

This thesis seeks to investigate an alternative of adding in area-based factors to "rescale" individual estimations of non-heating end-use energy of households instead of this rescaling process happening at the national level. There are several methods that are available to the researcher, all with different strengths and weaknesses, including multiple regression, multilevel regression, and ecological regression. Chapter 5 will outline the methodological options available and the consequences of each approach.

3.6 Fourth generation of BREDEM: becoming the second generation of SAP

3.6.1 Introduction

The fourth generation of the BREDEM model from 2005 is disseminated only through the publication of SAP. The Building Research Establishment was privatised in 1997, and therefore its publication of research became less frequent, and the last publication of the BREDEM algorithm was in 2001.

Therefore, any changes to rescaling factors in BREDEM's algorithm that might be used in stock modelling are unknown as these are not included in SAP. The last revision of SAP in 2010 introduced more complex ways for estimates of lighting to be reduced, and the introduction of assumptions about the effect of general weather conditions on the usage pattern of all non-heating end-uses. The conditional demand analysis model of non-heating end-use energy published in the Standard Assessment Procedure in 2010 is a very different type of model from the Uglow model of 1981 because of how it forms part of building regulations and the conditional demand of appliances' energy use formed as subtractions from a baseline level instead of summing up energy use per appliance owned.

3.6.2 SAP2005

In pre-2006 versions of the building regulations simplified versions of SAP calculations were allowed to be used (Office of the Deputy Prime Minister, 2000). Appliances, lighting, and metabolic gains were linked together in a single equation. Following the 2002 European Directive for the Energy Performance of Buildings (European Parliament and European Council, 2002), the 2006 building regulations incorporated Target and Dwelling Emission Rates in its version of Approved Document Part L (L1 for dwellings) (Department for Communities and Local Government, 2006a).

This migrated the parts of the BREDEM algorithm into SAP2005 for calculating both internal gains and electricity consumption due to lighting, but not appliances or cooking. SAP continued to be published by BRE on contract with the Secretary of State for housing and planning; the underlying BREDEM algorithms for use in conjunction with BREHOMES were no longer made public (Department for Communities and Local Government, 2006a, Office of the Deputy Prime Minister, 2000).

In this version of the Standard Assessment Procedure, a new appendix was published that explicitly calculates the electricity consumption of the dwelling due to lighting. This algorithm adopts a reduction in electricity use due to the installation of low-energy light fixtures. This algorithm adopted the Chapman function first presented in the Milton Keynes Energy Cost Index project, where up to half of the energy demand could be reduced with low-energy fixtures. A second category for reducing lighting demand was introduced – daylighting. Based on the orientation of the building and the amount and type of glazing in the building in windows and skylights, estimated electricity demand could again be reduced (BRE, 2005a). The presence of low-energy light bulbs and daylighting were expressed as correction factors in the equation

$$E_L = E_B \times TFA \times C_1 \times C_2 \quad (18)$$

where E_L is the estimated annual energy consumption of lighting, E_B is a constant baseline amount of annual energy consumption per square metre of usable floor area, TFA is the total usable floor area in square metres, C_1 is the correction factor due to low-energy light fittings (less than unity), and C_2 is the correction factor due to the proportion of daylighting based on the sum of light transmitted through windows in the building and floor area (BRE, 2005b).

The introduction of correction factors altered the unit energy consumption equation to a conditional demand analysis model

$$UEC_{ijt} = F_j(ST_i,\ AF_j,\ WC_{it}) \quad (19)$$

where ST_i is the size and amount of window glazing of household *i*, AF_j is the appliance features of devices *j* (proportion of low-energy lights only), and WC_{it} represents the amount of sun entering the household.

3.6.3 Current revision SAP2009 in place from 2010

SAP2009 was released for consultation in April 2009 and was released in its definitive version in March 2010 for implementation in building regulations in the United Kingdom from October 2010. SAP2009, like its predecessors, is a heat balance equation solving for the space heating requirement of a dwelling taking non-heating end-use energy as a given, labelled as "internal gains". However, the method of calculation of the energy consumption from which the internal gains are derived is explicitly stated in SAP for the first time. The area of lights and appliances has been, in the words of BRE, "an area in which there is limited information available and so it has been necessary to use data from the late 1990s (a follow-up survey to the 1996 English House Condition Survey) together with educated estimates" (BRE, 2009).

The changes to the model in SAP2009 in the area of appliances and lighting were threefold. The estimate for uses of appliances changed from a quadratic linear regression model based on a normal distribution to a growth model based on linear regression of a log-normal distribution. The model of electricity use changed from an annual to a monthly model.

A new algorithm was put in place, after conversations between the Department for Communities and Local Government and BRE determined that the modelling figures were too low (Lowe, 2010, BRE, 2009, Sefton and Chesshire, 2005). The new algorithm (Standard Assessment Procedure – 2009 edition) for appliances and lighting has been in place from October 2010. Because there are no consumption figures published for cooking, its fuel consumption needs to be estimated. The basic algorithm is as follows:

Lighting (kilowatt-hours per year):

$$E_L = 59.73 \times (TFA \times N)^{0.4714} \quad (20)$$

with the possibility of being diminished by low-energy lighting or daylighting correcting factors

Electrical Appliances (kilowatt-hours per year):

$$E_i = 207.8 \times (TFA \times N)^{0.4714} \quad (21)$$

then for month t (January = 1 to December = 12),

$$E_{A,t} = E_i \times \left[1 + 0.157 \times \cos\frac{2\pi\,(t-1.78)}{12}\right] \times \frac{n_t}{365} \quad (22)$$

where n_t is the number of days in month m

then sum all months to the annual total,

$$E_A = \sum_{t=1}^{12} E_{A,t} \quad (23)$$

Cooking (kilograms of carbon dioxide per year):

$$M_C = \frac{131+26N}{TFA} \quad (24)$$

The correction factors that were calculated for SAP2005 for low-energy lighting and daylighting were retained. Therefore, the unit energy consumption factors as part of the conditional demand analysis model remain unchanged, but the functions are altered because of seasonal variations between every month *t.* The evidence base for the seasonal fluctuations is derived from the change in surveyed electricity use by quarter in the 1996 English House Condition Survey and further evidence from National Grid substations (Coker, 2009), but surveys today does not have as solid an evidence base because of electricity market fragmentation into quarterly, monthly, prepaid, and other non-standard methods of payment (Smith, 2011).

3.7 The future of domestic energy modelling for England

The current revision of the domestic energy model for non-heating end-uses in individual buildings inevitably leads to changes in the assumptions and scaling factors in bottom-up housing stock models using BREDEM/SAP at the individual level. The latest change in domestic energy research in the UK is the transfer of responsibility for government-led research on housing stock modelling from BRE to Cambridge Architectural Research (CAR). This is leading to a change from BREHOMES to the Cambridge Housing Model. This is leading also to a revision of the factors in the algorithm of unit energy consumption of non-heating end-use energy (Cambridge Architectural Research Ltd., 2011).

Thinking about the role of non-heating end-use energy, beyond merely a source of internal heat gains in a heat balance equation is pushing non-heating further forward in the focus of energy researchers. The amount of non-heating energy consumption in residential buildings is set to rise dramatically as traditional appliances such as refrigerators and washing machines become more efficient, and on the other hand newer, more sophisticated electronic gadgets for communication and entertainment will mostly use more energy than the ones they replace. Although 54 per cent of people in the UK think that modern, high-tech kit is more energy efficient than older technology, the opposite can be often true (Energy Saving Trust, 2007b). From 2001 to 2020, entertainment, computers and gadgets are predicted to rise from 12 to 45 per cent of electricity used in our homes (Energy Saving Trust, 2007b).

The current research by CAR on *Great Britain's Housing Energy Fact File* (Department of Energy and Climate Change, 2011) involves building a bottom-up energy stock model using SAP, but not BREDEM, and testing the bottom-up model against actual energy use overall, measured by fuel consumption reported by Middle Layer Super Output Area (MLSOA) (Palmer, 2010). This bottom-up housing stock model will be called the Cambridge Housing Model and represents a break from BREHOMES in the development of housing stock models in the UK, although the model remains based on SAP for modelling the energy use of typical buildings or archetypes in a similar way as in the past (Shorrock and Dunster, 1997, Steemers and Cheng, 2010). The researchers involved in the preparation of the 2011 *Housing Energy Fact File* state that the successes of energy conservation the last 40 years were in water heating and cooking, but the weaknesses of policies were in space heating, appliances, and lighting, where demand continues to rise. They have also stated that they do not consider non-heating energy end-uses such as appliances and lighting to be part of SAP (Palmer, 2010).

The Cambridge modelling team is piloting an empirical model to attempt to "explain and predict building energy use patterns and the complex inter-relationships" between socioeconomic factors, physical characteristics of buildings, user-control behaviour of heating systems, and overall building energy consumption. There has been a pilot study cited from the United States, where the correlation between socioeconomic factors and floorspace, and income and heating energy use were explored previously (Energy Information Administration, 2001). This model could be used as a scaling factor to match the actual usage as measured by DECC (Steemers and Cheng, 2010). It does not propose a methodology that might insert these factors into SAP, nor has this work moved beyond heating energy end-uses and comparing energy-saving technologies targeted at space and water heating. It might also introduce the same weaknesses of group-level socioeconomic

assumptions and scaling factors into heating end-use energy that have already been explored in this thesis in the development of BREHOMES for rescaling of non-heating end-use energy.

This thesis proposes to explore both individual household-level and area-level variables, to avoid some of these same pitfalls that have been occurring in the development of bottom-up housing stock models based on conditional demand analysis models of non-heating end-use energy. The main barrier to researchers in the use of housing data is the lack of identification of individual cases. For example, CAR faces a significant barrier in that there is housing energy consumption data from sources such as the English House Condition Survey Fuel Sub-sample of individual households that indicates which MLSOA they are located in – the smallest spatial resolution is the nation or region. This thesis proposes to overcome these problems with the use of area. Area classification data on the English House Condition Survey has been obtained, and this is a unique dataset that has not been explored by researchers in energy and the built environment (McIntyre, 2011). The following section will explore the origins of area classification schemes for housing in England, and the basis for the latest classification system that is a national statistics for the United Kingdom with a methodology and dataset that can be interrogated by outside researchers (Vickers and Rees, 2007).

3.8 Area Classification

3.8.1 Introduction

The area classification system that is now defined as a 'national statistic' by the Office for National Statistics was developed in conjunction with the University of Leeds in the mid-1990s. This section will trace the origins of this system from inception to adoption. First, we consider the principles of classification and cluster analysis, or placing groups into relatively homogeneous supergroupings. The history of this type of analysis in the context of the United Kingdom will be described alongside criticisms of the method and any conclusions that are subsequently drawn. The reasons for choosing the 'national statistic' for data matching instead of commercial alternatives will be explained. The methodology of classification by the Office of National Statistics will be described, alongside strengths and limitations pertinent to their application to the examination of non-heating end-use energy. Finally, the selection of variables for classification will be described, centring on the built environment and socioeconomic characteristics identified previously. The conclusion of this section on area classification is a set of groupings that are predicted to have significantly different non-heating end-use energy consumption in households for testing later in this thesis.

3.8.2 Principles of classification and clustering

Area classification is a method of defining geographic patterns from a range of variables by identifying similarities and dissimilarities between areas (Webber and Craig, 1976). The method is a categorisation and not discretisation of these patterns. This means that nouns, not numbers are used to describe, or label, the differences between categories with no definable numerical interval between categories. Everitt (2011) defines classification schemes as "a convenient technique for the organisation of a large dataset to enhance the efficiency of information recovery." The use of labels produces a summary of the arrangement of differences and similarities between objects being described in the data.

Classification theory can be defined as "learning by similarity." This concept recognises that no two objects are completely identical, but the human mind increases its understanding of all objects by grouping similar items together. By forming these groups, people know how to react to and what to do with a new object when they have built up a knowledge base of similar objects (Pinker, 1997). Classic examples of formal classification can be found in chemistry, with the original justification by Mendeleev for the groups in the table based on recurring trends (International Union of Pure and Applied Chemistry, 2010), and in taxonomy, with the splitting of vertebrates and invertebrates by Aristotle in order to aid understanding of the underlying properties of organs and animals (Everitt, 2011).

Vickers, in his work that underpinned the area classification system now used by the Office for National Statistics, defined his classification of residential areas on the principle of Carl Linnaeus's *Genera Plantarium* of "distinguishing the similar from the dissimilar" (Vickers, 2006). The argument made by classifiers of the areas where people live is that the complexities and statistical noise produced by the aggregation of people into small areas (there were 32,482 Lower Layer Super Output Areas and 223,060 Output Areas defined by the Office for National Statistics after the 2001 Census) is too great for any human understanding and therefore classification is a necessity. Classification reduces the amount of data and noise to a point where researchers can see patterns in the distribution of area typologies, and can start gathering information on what processes are taking place in society and technology.

3.8.3 History and criticisms of area classification in the United Kingdom

This section will describe the development and history of area classifications of the residential population from the classic work of Charles Booth in the United Kingdom and the "ecological" or Chicago School in the United States to the present-day geodemographics and marketing industries. Area classifications are discovered in research by summarising the profiles of areas or uncovering

clusters of similar areas. These approaches both lead to the same outcome of the creation a categorisation system for geographically defined areas.

Area classification broadly traces its roots to the work of Charles Booth in London (Davies, 1978). His detailed work in *Life and Labour of the People in London* published from 1889, categorised all of the streets of London into one of seven social classes. Booth's work originated as a response to sample surveys carried out in London by a liberal organisation called the Social Democratic Foundation who had concluded that around a quarter of London's residents were living below the poverty line. Instead, his area classification system estimated that the figure was closer to one-third of the population, which went against his hypothesis that 25 percent was too high a figure to be believable (Booth, 1902).

Booth's study was one of the first examples of use additional datasets with his own field research. He started from large-scale, census-like data obtained from the School Board for London, which collected data on the housing conditions of every household that contained school-age children, and from the vital records of births, deaths, and marriages from London's parish churches. He then hired a team of researchers to pioneer survey and interview techniques to sample London's population, to add more variables in his investigation of poverty (Simey, 1960). This led to a composite classification of social condition based on six variables selected from a longlist. These were poverty, overcrowding in dwellings, early marriages, unmarried young and middle-aged adults, the birth rate, and the death rate. Area classification methodology originated from his ranking of areas without regard to exact intervals, the use of census-derived variables with survey data, combining and correlating several variables, but with the weaknesses of correlating these same variables within the same household, instead of describing the correlations in areas (Davies, 1978).

Finally, Booth also generalised his cartographic representations by classifying groups of houses from street corner to street corner, or of an individual mews or alley, instead of mapping the differences between houses in an area. There were real differences between homes in an area, but attempting to map out all of these differences would not have illustrated the location of social classes throughout London (Harris et al., 2005). The strengths and weaknesses of area classification derives themselves from these self-imposed limits on explaining reality.

This type of area classification of social conditions is different from the household classification system for the United Kingdom created by the Registar General after the 1911 census and used until the introduction of the National Statistics Socio-economic Classification System (NS-SEC) after the 2001 census (Rose and Pevalin, 2003). This classification system was created from information

provided by the (male) head of household on his employment or occupation in the census, to investigate the correlation between occupation and infant mortality (Haines, 1995). It has been widely taught in social studies classes in school for the last 100 years and is widely understood by the majority of the UK population. The categories are:

A) Professional occupations
B) Managerial and technical occupations
C) Skilled occupations
    1) Non-manual
    2) Manual
D) Partly-skilled occupations
E) Unskilled occupations

This method of determining social class by the numbers of heads of households within each social class per census area, however, is not area classification. Several decades would pass before similar work was done on the selection and correlation of variables in the same manner as Booth's studies of London at the turn of the century.

Urban planners and sociologists working in the United Kingdom in the post-war period, instead of looking to Booth, looked to the methodologies developed by the "ecological" or Chicago School of urban researchers based at the University of Chicago in the early 20th Century (Herbert, 1972, Johnston, 1971). The "Chicago School" developed representations of cities in the context of Chicago that were subsequently applied to other American and western cities. The concentric ring model was developed by Ernest Burgess and distinguished five rings around the centre of a city (Lutters and Ackerman, 1996):

1. City Centre
2. Transition
3. "Workingmen's homes"
4. Residential
5. Commuters

This method is widely discussed and studied by urban planners and demographers. There were further theoretical models developed by the Chicago School such as Hoyt's sector model which divided the concentric rings by land use type based on set travel corridors, and Harris and Ullman's Multiple Nuclei model where land uses with complimentary uses are co-located using Ullman's

Gravity Model, and congestion-free travel is more dispersed and randomised (Robson, 1971, Harris, 1997).

This type of analysis did not include detailed census information on the aggregated totals of the entire population of a neighbourhood and instead relied on survey and interview data. However, by the 1960s, the United States Census started to make available aggregated data for urban areas by census 'tract'. These aggregate totals were investigated at the University of Chicago by classifying metropolitan areas using both statistical analysis and surveying of residential patterns across twelve major American cities within the context of the previous area classification models created by the Chicago School (Rees, 1979). These social area classifications were built on variables selected to explain social stratification, lifestyle choices, and the ethnic composition of American neighbourhoods (Timms, 1971). The area classification systems created by the Chicago School were created for the American context, but without sufficient research into classifications for European or non-western contexts, these models became partially entrenched within British government, local authority planning departments, and academia (Thornley, 1991). However, the application of these classifications at the national level was fraught with difficulties because of their origins in case studies.

A decade later, small area statistics began to be produced for Great Britain which led to the commissioning of national classification systems during the late 1970s by the Office of Population Censuses and Surveys (OPCS) (Webber and Craig, 1976). 40 variables from the 1971 census were selected to classify wards and parishes throughout Britain to place them into 36 clusters; then enumeration districts were placed into 53 clusters. The variables selected included employment levels, car ownership, employment sector of heads of households, age, immigration and migration, housing tenure, overcrowding, and household amenities such as bathrooms, kitchens, and central heating (Webber, 1978).

After their release, these classifications began to interest market researchers for studying consumer patterns instead of conducting sample surveys. The national classification of wards and parishes was then renamed A Classification of Residential Neighbourhoods, or ACORN and launched at the 1979 national conference of the Market Research Society (Beaumont and Inglis, 1989). After the 1981 census, other classification systems were created and by the end of the 1980s four major commercial classification packages were created by the market research industry – ACORN, PiN, Mosaic and Super Profiles (Vickers, 2006).

Strident criticism came from the academic community warning local authorities and other public and private decision-making bodies after over 50 local authorities bought access to the classification scheme only months after the release of the system by the OPCS. Openshaw et al. (1980) stated that the creation of this new methodology "should have been seen as a pioneering first attempt rather than as a polished final product based on the application of well established, proven methods." The paper criticised the pre-cursor of ACORN for lacking a single objective in the classification methods, and the fact that the creators of classification methods were justifying their use as there were better methods in academic discourse at the time, and the frequent use of the classifications by decision-makers. This use by decision-makers was making regeneration funding streams available by targeting areas, not households, identified by area classification. This concern led to further work leading to Openshaw's paper on ecological regression and the Multiple Areal Unit Problem in 1984 (Openshaw, 1984). Ecological regression will be further discussed in Chapter 4.

Commercial area classification, or segmentation, packages used additional data on top of the publicly available census data. The original national classification of wards used only census data, but the commercial systems added new information that was focused on identifying affluence in society (Harris et al., 2005). The addition of non-census data sources was intentionally built into the original classification procedure (Webber, 1978). The classification agencies would mine additional information, including sensitive personal information that would come into the control of parent companies involved with credit scoring and county court judgements involving financial judgements and defaults. Mosaic, for example is part of Experian, a large credit scoring agency. Mosaic also has purchased electoral roll records and vehicle registration data, then taken additional sample surveys on lifestyle choices as well as appliance warranty records that are commercially available (Experian UK, 2009).

However, there are major criticisms of the bias created by the use of this data, namely that it does not cover the entire population whilst being opportunistic, and not systematic, when collecting additional data. The bias is towards collecting information on non-poor, non-immigrant backgrounds and gives a picture of a more affluent nation (Vickers, 2006). The raw data is not publicly available, and therefore cannot be verified by external researchers.

3.8.4 Choice of the national statistic for area classification

A new, general-purpose area classification system has been made on behalf of the Office for National Statistics that can be re-examined by external researchers since 1996, with the current system adopted in 2006 (Vickers and Rees, 2007). This classification system is currently known as the National Statistics 2001 Output Area Classification. The system "places each output area in a group

with those other output areas that are most similar in terms of census variables" (Vickers, 2005). The same system is used for Lower Layer Super Output Areas, at the same level of spatial resolution of energy consumption data released by the Department for Energy and Climate Change (Department for Energy and Climate Change, 2010b). As the classification procedure and its methodology is open for scrutiny, it is a valuable area classification system for use in this thesis as opposed to commercially generated alternatives.

This is the second classification system produced by the Office for National Statistics. The 1991 Area Classification was released in 1996 as a classification of local and health authorities throughout the UK. However, these did not have the same spatial resolution as commercial products which were built on enumeration districts (around 200 households). One example is the use of ACORN groups to evaluate which areas were at greatest domestic fire risk (Vickers, 2006).

The 2001 Output Area Classification applied to three geographic layers for Britain. The smallest is the output area, which contains around 40 households. The next level up on the geographical hierarchy, also known as the ONS hierarchy, is the lower layer super output area (LLSOA) of around 400 households, and then the middle layer super output area (MLSOA) of around 2,000 households (Martin, 2001). The area classification principles are repeated for each level of the ONS hierarchy using the same methodology developed for output areas by the research team at the University of Leeds. Area classification at larger scales such as the local authority or ward level are calculated using a separate methodology (Office for National Statistics, 2008b).

The 2001 Output Area Classification was created with census data and without any outside proprietary data sources. Raw data collected by the census and released in aggregate is freely available for researchers to re-create, evaluate, and critique the methodology, notably in the selection of variables and labels. The methodology is also freely available, and will be detailed in the following section. This thesis proposes to use the 2001 Output Area Classification in its research into the combination of individual household-level and area-level variables to predict the amount of non-heating end-use energy in households.

3.8.5 Output Area Classification Methodology

The area classifications for output areas in England were constructed by creating a hierarchy of clusters, or groups, which characterise different areas. This continues the practice of creating hierarchies of classifications since classifications made by the OCPS in the late 1970s for wards. There are seven supergroups, 20 groups, and 53 subgroups in the 2001 Output Area Classification. This section summarises the classification and its methodology, though more detailed notes can be

found in the University of Leeds published by the Office for National Statistics (Vickers, 2006, Vickers and Rees, 2007, Office for National Statistics, 2008b).

The supergroups and groups are labelled by examining the average socioeconomic and environmental features of the clusters. The methodology makes it clear that it is possible in any cluster to find that there are outlying super output areas that are not accurately described by the label of the cluster. Table 2 below details the final division of supergroups, groups, and subgroups:

Table 3.12: Cluster labels and hierarchy

Super ID	Supergroup Name	Group ID	Group Name	Sub-groups
1	Countryside	1.1	Countryside communities	a,b,c
1	Countryside	1.2	Rural economies	a,b
1	Countryside	1.3	Farming and forestry	a,b,c,d
2	Professional city life	2.1	Educational centres	a,b
2	Professional city life	2.2	Young city professionals	a,b
2	Professional city life	2.3	Mature city professionals	a,b,c,d
3	Urban fringe	3.1	Urban commuter	a,b
3	Urban fringe	3.2	Affluent urban commuter	a,b
4	White collar urban	4.1	Well off mature households	a,b,c
4	White collar urban	4.2	Young urban families	a,b
4	White collar urban	4.3	Mature urban households	a,b,c
5	Multicultural city life	5.1	Multicultural inner city	a,b,c
5	Multicultural city life	5.2	Multicultural urban	a,b
5	Multicultural city life	5.3	Multicultural suburbia	a,b,c
6	Disadvantaged urban communities	6.1	Struggling urban families	a,b
6	Disadvantaged urban communities	6.2	Blue collar urban families	a,b
7	Miscellaneous built up areas	7.1	Suburbia	a,b,c,d
7	Miscellaneous built up areas	7.2	Resorts and retirement	a,b
7	Miscellaneous built up areas	7.3	Urban terracing	a,b,c,d
7	Miscellaneous built up areas	7.4	Small town communities	a,b

Some classifications are found more frequently in different regions, whereas some are equally spread throughout the country. Supergroups that label characteristics that typify some urban lifestyles, such as professional city life and multicultural city life, are concentrated around major cities with their urban fringes. Other supergroup labels that concern urban lifestyles, such as white collar urban, disadvantaged urban communities and miscellaneous built up areas, are found in towns and cities up and down Great Britain. The countryside supergroup, although not covering most of the population, covers most of the land area of the UK.

The supergroups display characteristics that can, on average throughout the cluster, be above, below, or the same as the national average. For example, a super output area labelled in the urban fringe supergroup will typically have less terraced houses and flats, higher proportions of homes with central heating, and lower proportions of social housing than the national average. A super output area labelled under the white collar urban supergroup is typified by its “averageness,” as the characteristics of age structure, household size, and occupation that distinguish this supergroup are average.

An example cluster summary as a radial plot for the output area supergroup “countryside” is given in Figure 3.2. (Note: supergroups at output area and super output area levels are different and non-exchangable.)

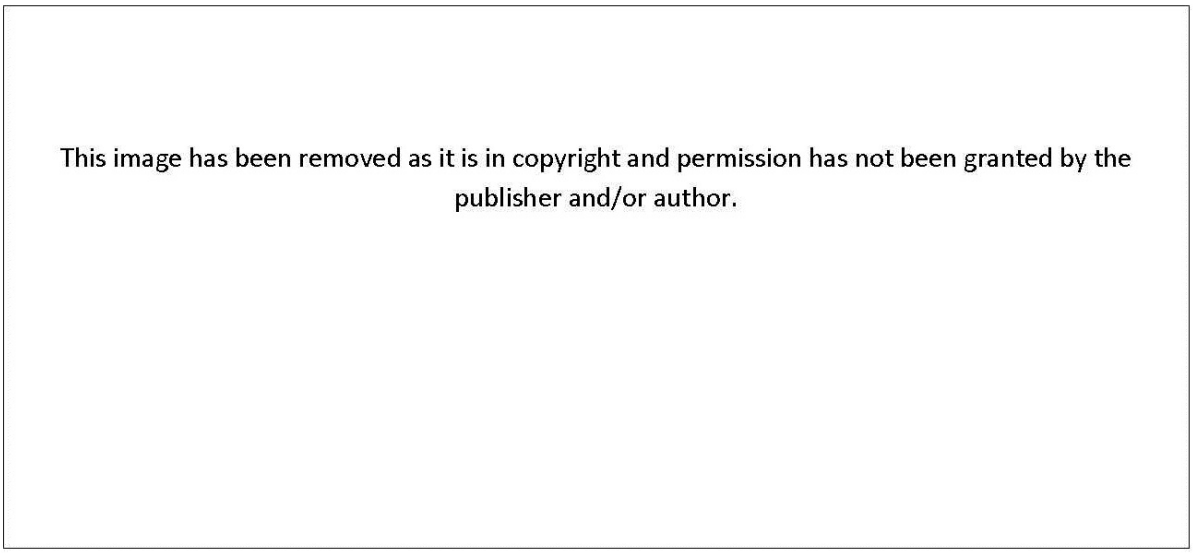


Figure 3.2: Radial plot for Supergroup 3 "countryside" in Vickers (2006)

were rescaled to the number of standard deviations above or below their means, to give each variable equal weight. The decision to give each variable the same weight was made because the classification is intended for general purpose use. The argument was made in the development of the system that there are enough inter-correlations between variables in the data that adding weightings will have unpredictable effects on the final classifications (Vickers, 2006, Vickers and Rees, 2007). This problem of inter-correlation of both socioeconomic and built environment variables has also been reported in several publications the interactions of humans and energy use, especially for electronics, lighting, and appliances (Lutzenhiser, 1992, Wall and Crosbie, 2009, Boardman, 1988). Therefore it is wise not to attempt to improve the classification by rerunning the methodology, weighting some variables higher than others.

Variable choice was driven by the principle of selection of the "minimum number of variables that satisfactorily represent the main dimensions of the 2001 Census." (Vickers, 2006) The variables were chosen only from the census because of its comprehensiveness and the re-distribution of aggregate data collected at different scales from those within the ONS hierarchy. Five main domains of variables were chosen, labelled Demographic Structure, Household Composition, Housing [stock], Socio-Economic, and Employment, from the variables previously classed by the ONS as their Key Statistics (Office for National Statistics, 2004). Highly correlated variables that signalled redundancy were identified, eliminating some variables from consideration. Highly skewed variables or variables with large numbers of null values for an output area were also eliminated. Other variables were eliminated on the grounds that the questions were only asked in certain parts of the UK, the questions were vague, or the questions addressed issues that had factually changed or will change in the lifespan of the classification. Chapter 5 will details the differences between the variables in the classification and variables that are have previously been identified as correlated with non-heating end-use energy in the home, and how they are documented within the 2001 Area Classification for output areas and super output areas.

The clustering method used was a repeated application of the *k-means* method to create a hierarchical classification system. The *k-means* algorithm is an iterative procedure designed to minimise within-cluster variability using the sum of squares as the error or variance measure. *K-means* moves an area from one cluster to another, to examine if this action improves the sum of squared deviations within all the clusters. This process was repeated for every area and cluster until a stable classification was reached. After the clustering was complete, the means of each cluster were compared for each variable and domain, to ascertain the distinctive qualities of each cluster (Vickers and Rees, 2007). The minimising of variance for validation, or the minimisation of the percentage of within-cluster sums of squares, has been a focus for geography researchers since their introduction, in the criticisms of the OPCS ward classification in the late 1970s (Openshaw et al., 1980).

The "sums of squares", representing dissimilarities, or distance, is represented below:

$$D_c{}^2 = \sum_{i=1}^{n_c} \sum_{j=1}^{m} (Z_{ijc} - \bar{Z}_{jc})^2 \qquad (25)$$

Adapted from (Vickers and Rees, 2007)

where $D_c{}^2$ is the squared distance between the characteristics of all areas in cluster *c* representing dissimilarity, n_c is the number of areas in the cluster, *m* is the number of variables, Z_{ij} is the value of variable *j* for area *i* in cluster *c* with $\bar{Z}_{cj}$ as the cluster mean for variable *j.*

The numbers of clusters at each level of the hierarchy were taken as *a priori* targets for the number of clusters following consultation of potential users of the classification system. These targets were suggested in order to optimise the visualisation of supergroups, customer profiling in groups, and market propensity in subgroups. Labels for the groups and supergroups were designed to not contradict other official classifications or duplicate names in use by commercial competitors, and after public consultation on the initial labels proposed. (Vickers and Rees, 2007).

The use of a general use classification system is good for the evaluation of area-level variables in non-heating end-use energy. The spread of socioeconomic and built environment factors in determining the classifications is broadly reflective of the spread of models of human behaviour that contribute to non-heating end-use energy and that will be covered in detail in Chapter 4. The system also does not include factors that are worded exactly in the same way as in the SAP model, using rooms per household and people per room instead of the floorspace per household and people per household. Therefore it is not required, and perhaps unwise, to eliminate them as additive effects in area-level variables already present at the individual level. The use of the classification as classes summarising area-level variables will be explored further in Chapter 5.

3.9 Conclusions and research directions to pursue

This section described the methods, results, and criticisms of previous work in domestic energy modelling of non-heating end-use energy and area classifications in the context of the United Kingdom. The overall conclusion is that the correct research to pursue is to explore the combination of individual and area based variables using classifications as area-level groups. Details about the socioeconomic status of a household have been eliminated from domestic energy models adopted by government in England along with attempts to estimate the ownership of electronic devices in a household. However, the feedback from using individual level models to build up a housing stock model shows that the elimination of these factors was unwise, and these were retained through the "back door" of housing stock modelling. Area classification has been introduced as a way of introducing these factors at a smaller spatial scale than the national level. This research will therefore focus on the role of area-based variation in socioeconomic and built environment factors in the modelling of domestic non-heating end-use energy in single households.

From this survey of work relevant to the housing context of England, this research will compare the correlations of non-heating end-use energy with socioeconomic and built environment factors. Previous research and models have shifted back and forth between socioeconomic variables such as appliance ownership, and built environment variables, notably usable floor area.

Socioeconomic factors at the area or aggregate level should be investigated instead of the individual household level. There is a clear message from policy makers that estimating the social class, income, or other status of an individual householder is unacceptable. In its place, this thesis will explore area-level socioeconomic variables whilst respecting data protection rules that will be detailed more fully in Chapter 4. In order to simplify the thousands of aggregates of non-heating energy use available from small area statistics, area classification measures are a logical way of organising and evaluating these factors.

The following chapter will introduce the data sources available to the researcher for investigating the non-heating end-use energy of households in the context of England. It will begin by explaining the types of model of human behaviour and interaction with the use of energy in housing and what data is appropriate to conventional and unconventional perspectives. Data will be introduced, assessed, and rated for its appropriateness in conducting the analysis that uses both individual and area-level variables for estimating this kind of energy use in buildings.

Chapter 4 - Data sources

4.1 Introduction

This chapter considers issues associated with the measurement of non-heating end-use energy at the individual household and neighbourhood level in the UK.

This chapter is divided into five sections. The first summarises the different approaches in academia and governments of what data should be used to create a model of non-heating end-use energy, especially in the context of the United Kingdom. The second section describes criticisms of the different types of data available that could measure end-use energy. The third illustrates the methodological difficulties in collecting data and sources of inconsistency in defining and measuring domestic energy. The fourth lists the specific data sources that are available to researchers at the national level for England. Finally, the fifth summarises the findings of an analysis of data sources.

4.2 Models of household energy consumption

4.2.1 Types of models

The approaches for determining causal relationships in the energy use of a household are contested, creating different workstreams for academics to approach the same problem of measuring, then predicting, household end-use energy (Guy, 2006). The fundamental question of what is known about household energy use has been asked time and again, and almost always one of the following conclusions is arrived at, at the end of a study:

- From the perspective of our discipline, households perform in the way demonstrated by the model (Crosby, 2006), or
- Because of the fundamental disconnects between the interests, theories, methods and conclusions of the disciplines concerned, the only certainty is uncertainty (Keirstead, 2006).

Four broad models have been developed to conceptualise non-heating end-use energy as part of the measurement of single household energy performance : physical-technical-economic models, economic models, psychological models, and sociological models. The perspectives, theories, and data sources required by each of these models differ considerably from one another. Each of the models looks at the end-use consumption of energy in the household in a different fashion,

emphasising the data that play to the researchers' strengths while making assumptions or ignoring data that lends itself to analysis using the methodologies of other disciplines. (Stern, 1986) described this as an intentional "blind spot" using economics as an illustrative example where researchers will simply put the data outside their interest into a "black box" which holds it as a constant, explain it as a sequence of random events, or simply place it outside the frame of research. Lutzenhiser et al. (2010) went further than this, declaring that "the omissions are unacknowledged and disciplinary analysts may be so 'frame-bound' as to not even be aware of them. So when we attempt to look at the literature, we find a Tower of Babel of disconnected strands and compartmentalized theories." Therefore the choice of a data framework that will enable both individual level and area level predictors is extremely important, as some modelling frameworks interpret them under their own disciplinary bias.

The modelling framework that is most often used in formulating energy policy by governments is what Lutzenhiser, Sullivan, and Guy have described as either a physical-technical-economic model or a building science-economic science model. This model assumes that energy-efficient choices emerge out of the improved awareness of technologically superior alternatives, and that the cost of these alternatives comes down to a level that enables people to make the 'right' and rational decision to pursue these alternatives. The data needed to support the development of such models is a list of devices that consume energy, from boilers to appliances, and the details of the building envelope beyond the floorspace and occupancy to include building materials and the architectural design of the building. Data to support this model is collected by trained building services assessors who can identify a variety of technical devices, building materials, and can measure a building footprint in a standardised fashion and estimate the cost of building improvements alongside a collection via interview of the socioeconomic characteristics of the household (Department for Communities and Local Government, 2010a). This survey methodology can be replicated when assessing buildings for projects that fund physical interventions such as the Warm Front project in England (Hong et al., 2006b). Examples of such models are found in the finest-grain simulation models of buildings that are in place as part of building regulations, for example the Standard Assessment Procedure in England (BRE, 2010).

Economic models attempt to estimate household consumption as a function of market price and price elasticity (Lavin et al., 2011, Espey et al., 1997, Bijmolt et al., 2005). Price elasticity is a quantitative measure of the change in energy consumption in response to a change in price. This is built on an overarching argument that either end-use demands will be reduced (fewer devices) or that efficiency will be improved (reducing the demand of the same device). Recent evidence in the

UK has show that there is some elasticity observable when controlling for weather in aggregated data (Summerfield et al., 2010). Data that is required for this type of modelling is not as detailed as the physical-technical-economic model, but prices of energy must be included with basic socio-economic data and must be longitudinal in order to conduct the analysis, as price elasticity is a relative measure.

Psychological models that deal with energy use and the built environment fall into what is alternatively called Attitude-Behaviour-Context or external conditions models (Stern, 2000). In this type of model, the attitudes and behaviours of the energy use of an individual or household are correlated with what an individual understands to be the social "norm". The types of data needed for psychological models are varied, but are focused on quantitative data for experiments that harmonise with the physical-technical-economic model (Lutzenhiser et al., 2010). Those that focus on the choices and behaviours of individual households require housing survey data that includes the socioeconomic and built characteristics of the household and house, but also a measure of the view of the social norm for each individual household. Some experiments also require a longitudinal dimension in order to measure the different levels of self-awareness of social norms, such as knowledge of the energy usage of their peers (Schultz et al., 2007).

Sociological models also concern themselves with the social norms of end-use energy in the built environment, but do not explicitly set themselves up as representing a choice feedback mechanism for physical-technical-economic model in the same way as in psychological models (O'Neill and Chen, 2002). They are interested instead in the way that household actions are patterned, shared amongst and between groups, and shaped by social, cultural, political, and economic trends. They study the context in which the household belongs, with the view that any decisions of the individual household, including how much energy to use, are a part of much larger patterns and practices. The data for these models requires that household survey data must include an indication of membership of the household to identifiable groups that connects into a classification and/or segmentation of the population relating to the built environment as well as socioeconomic variables.

These models are not mutually exclusive and there is extensive crossover between social science disciplines and building science-led physical-technical-economic model in feedback and experimental mechanisms. One example of this is the re-introduction of passive ventilation, otherwise known as openable windows, in office buildings after occupants were "irrationally" opening windows in hot weather and overloading cooling systems despite the calculation of engineers that assumed a sealed building envelope (Jelsma, 2002).

4.2.2 Stratified sampling

The different perspectives of the models do impact upon the questions used to measure end-use energy in surveys, especially the choice of strata in the design of large-scale household surveys. The major indication that can be used to determine this predisposition of model type is the grouping of respondents for analysis, known as stratified sampling. Typically, with stratified sampling, segments of the population are intentionally over- or underrepresented by the sampling scheme. There are two common types of stratified sampling, standard stratified sampling and variable probability sampling. Variable probability sampling is a method of sampling if there is no data on the respondents ahead of the survey. Therefore, standard sampling techniques are used by government-led surveys because they have a great deal of information on the respondents due to access to geolocated census data not available to researchers in academia.

Standard stratified sampling, also referred to as stratified random sampling, is the partitioning of an entire population *Y* into *G* non-overlapping groups $\{Y_g: g = 1,2,3, \dots G\}$. A random sample is taken from each sample stratum *g* $\{y_{gi}: i = 1,2,3, \dots, N_b\}$ where N_g is the number of observations drawn from stratum *g* and $N = N_1 + N_2 + N_3 + \cdots + N_g$ is the total number of observations. The assumption that is made to enable analysis of a stratified sample is that the distribution of the sample y_{gi} taken of each group *g* is the same as the distribution of the sample y that represents population *Y* provided that *y* belongs to group Y_g*:*

$$D(y_{gi}) = D(y|y \epsilon Y_g) \quad (26)$$

Table 4.1: Index of variables in stratified random sampling

Population	*Y*	Representative sample	*y*
Number of groups	*G*	Stratum	*g*
Population by stratum	Y_g	Sample of individuals by stratum	y_{gi}
Mean of the sample of the final stratum (equal to the number of groups *G*)	$\bar{y}_G$	Mean of the sample by stratum	$\bar{y}_g$
		Aggregate share by stratum	π_g

In order to estimate a central tendency for the entire population, for example, the mean of *y* from a standard stratified sample, the mean must be constructed using the aggregate shares of the strata. An aggregate share π_g is the probability of *y* being a member of group Y_g. Therefore, the estimate

$\bar{y}_g$ of the sample's mean of stratum *g* can be used with the aggregate share to estimate the overall mean of the population $\hat{u}_y$:

$$\hat{u}_y = \pi_1\bar{y}_1 + \pi_2\bar{y}_2 + \cdots + \pi_G\bar{y}_G \quad (27)$$

The optimum way of choosing strata could be different for each outcome variable, so the approach inside the government departments responsible for housing, for example, has been to define a small set of important outcome variables that are pertinent to the built environment, and to look for strata that perform well in reducing the within-stratum variance of most or all of these important variables. This key set of chosen variables tends to vary from one survey to another depending on the focus of the survey, and the same variable may even be used in different forms. Certain strata, such as social class measures, tend to recur in many household surveys, but it is not safe to generalise from one survey to another which strata are appropriate (Department of the Environment Transport and the Regions, 2002).

4.2.3 Selection of strata in surveys impact on selection of model

UK surveys on housing that include questions on household energy consumption have been designed on the whole to accommodate the physical-technical-economic model of energy use. For example, the strata chosen by the English House Condition Survey 1996 are made with the sampling stratified by year built, tenure (e.g. owner-occupied, social housing), type (e.g. detached house, converted flat) and region (Department of the Environment Transport and the Regions, 2000b). The fuel sub-sample of the English House Condition Survey 1996 surveys the metered fuel use of the household, and was set up with extensive input from the UK Building Research Establishment. It surveys, for example, predictors of energy consumption such as the improvements to the materials and insulation to the house by asking about the expenditure and value of the improvements, therefore obtaining data on the decision-making process as one built around the knowledge, availability, and affordability of technological advances in energy efficiency in heating, but not non-heating, end-uses.

The collection of general household survey data, such as the Living Costs and Food Survey, views energy use as a module of household expenditure, and therefore asks questions that are more appropriate for an economic or a sociological model of end-use energy (Office for National Statistics and Department for Environment Food and Rural Affairs, 2010). There are several studies in the realm of the psychological model of energy use that are broadly designed as intervention studies for the range of technologies that can be substituted for end-uses, and therefore designed to interface

with the physical-technical-economic model of end-use energy (Abrahamse and Steg, 2009). Again, this proves that most, if not all, models are hybrids.

The creation of strata by area classification has been an innovation that has been spearheaded by market research, and has started to gain favour in the academic research community, notably in the investigation of contextual effects in epidemiology (Webber, 1989, Chishimba et al., 2009). There are three main classification systems in use in the UK: Acorn, Mosaic, and the Office for National Statistics Area Classifications. The ACORN and Mosaic area classifications are designed as market research tools based on a great deal of economic data not available in the public domain, such as the Mosaic classification's ability to source credit scoring data from its parent company Experian (Experian UK, 2010). The ONS classifications, in contrast, are created as an alternative to the older economically-based classifications to input into a variety of sociological models by using socioeconomic and built environment strata (Howick, 2004). They are also available to researchers as data that can be attached to individual records, such as housing surveys, when data protection laws prohibit disclosure of the address of the survey participant.

However, they have not been previously used as strata by general household surveys for the following reasons:

- The labelling of area clusters is designed to "capture" many different dimensions of variation between areas and their residents.
- It can make areas that belong to different clusters appear to be more sharply distinguished from each other than is actually the case.
- The clustering algorithm is not optimally focused on any one area of application, such as housing (Department of the Environment Transport and the Regions, 2002).

4.3 Criticisms of data available

This section will review prevalent criticisms of data that is available to measure household non-heating end-use energy, and some counter-arguments to these criticisms. Researchers ideally want to have access to actual measurements of electricity and natural gas use, especially in the cases where these fuels can represent the split between heating and non-heating end-use energy in the household. Often, carbon-saving policies are measured against the number of installations in the household instead of being measured against actual energy demand reduction. Energy surveys that can be used for longitudinal analysis are either dated, or collection methods are so varied between projects that data harmonisation is difficult. Other types of data may be available, but they come with restrictions that limit their usability, such as partial identifying information, that limits the

ability of the researcher to correlate the relationship between the location of households and their energy use. Finally, data is often only available in aggregate, instead of at the household level. These criticisms are explored in the types of data that have been developed to accommodate the models of household energy use explored earlier, with particular attention to the physical-technical-economic model.

4.3.1 Introduction – focus on the physical-technical-economic model (PTEM)

The physical-technical-economic model of household energy use is dominant in the way that data is collected on households to support technical single-building simulations, such as BREDEM in the UK, HOTCAN in Canada, and PHPP in Germany. The first and major criticism is the focus on the measurement of the number of energy efficiency measures in the household and the modelling of the occupant as a physical construct instead of a social one, therefore rendering the model mostly unable to estimate the 'rebound' effect of efficiency measures apart from some upward adjustment of internal temperature. Other criticisms are that data collection comes from unrepresentative samples from 'captive audiences' in demonstration projects or social housing projects, and that data collection is the result of experiments designed with consumption data as a secondary result.

A technical, or engineering, perspective views non-heating end-use energy as the function of the size of the household (BRE, 2010, BRE, 2005b) or the sum of the electricity consumption of non-heating devices in the household (Yao and Steemers, 2005, U.S. Energy Information Administration, 1997, Parti and Parti, 1980). The advantage in the technical model is that the impact of technological change in devices or property trends in the size of households can be estimated. One disadvantage of this model is expressed by Lutzenhiser (1992) is that it "impute[s] self-conscious rationality to energy-use... of a level of intentionality that human action rarely possesses."

There are counter-arguments to these criticisms. The primary one is that built environment researchers should be able to write about the ability of policy to proactively reduce electricity demand and therefore carbon emissions, and that access to the right technology is the most realistic way forward for any government serious about meeting its targets on reducing emissions. The UK is very reluctant to use the price of energy to constrain demand. The second is that electricity will be decarbonised as a result of government-led electricity generation policies. Another counter-criticism is that data from small samples is valuable because in these experiments end-use monitoring is possible in a way that is not affordable using highly developed statistical sampling and data collection at the national level. The conclusion of this discussion is that the physical-technical-economic model of household energy use is supported by a large base of data that rests on technological advances in reducing heating, but not non-heating end-use energy.

4.3.2 Focus on recording installations, not measured consumption

The measurement of the number of energy-saving measures in dwellings has been a focus of the development of government-sponsored datasets in the UK. One example of this is the Homes Energy Efficiency Database (HEED), developed by the Energy Saving Trust (Neffendorf et al., 2009, Energy Saving Trust, 2008). HEED is a housing "survey of surveys" that collects datasets on energy efficiency and microgeneration installations – examples given include cavity wall insulation and solar hot water. The criticism of these databases is that the installations of heating and lighting because the results of these efficiencies can be explicitly determined.

In more recent years, the data on energy efficiency installations in HEED has since been linked to gas and electricity consumption at the dwelling level. As HEED links each dwelling to its gas and electricity bills, it is possible to look at the effect of energy efficiency measures before and after installation. However, these interventions documented all effect either space or water heating, and no electrical appliances are documented. The links to individual addresses were made available, on a contract basis to researchers working as consultants to the UK Department for Energy and Climate Change. A preliminary report by Bruhns et al. (2011) used this address data to find comparable groups within the entire stock by comparing those with or without a given measure. However, the data and its confidentiality agreements are not open to general scrutiny by academics and not available for this piece of research.

Between 1996 and 2011, the national survey of the housing stock did not measure actual energy consumption, it only estimated the energy use of a dwelling based on its characteristics (Department for Communities and Local Government, 2010a). The 2011 Survey of English Housing will include a sub-set of measured energy consumption, but funding for future measures of consumption are not guaranteed in future national housing surveys (McIntyre, 2011).

This problem can also happen in well-designed studies on energy efficiency installations due to low response rates from certain sections of the population. A study was recently conducted in the UK interviewing householders about their energy-saving attitudes to energy and water efficiency. The study found that owner-occupiers, heads of households over 40, and those with high household incomes were most likely to have invested in energy efficiency to their dwellings in the last five years (Department for the Environment Food and Rural Affairs, 2007). Unfortunately, there is no information available from the marketing company hired by DEFRA to conduct the interviews about the households who were approached and did not choose to take part in the survey.

The focus on modelled, as opposed to measured energy use in households when assessing energy performance in the UK context (Day et al., 2007) is limiting the ability of the PTEM model to measure

the ‘rebound’ effect of installing energy efficient installations. This criticism is known as either Jevons’s paradox or the Khazzoom-Brookes postulate relating to the proposed correlation between efficiency increases and consumption increases (Jevons, 1865, Sanders, 1992). In the words of Jevons, “[energy in the form of coal] is only saved from one use to be employed in others.” Full reviews of the arguments for and against Jevons’s paradox in the context of energy efficiency and conservation have been conducted recently (Sorrell, 2009, Alcott, 2005). This is important in the context of non-heating end-use energy as demand for electric heating is subject to a number of uncertain factors as the generation of energy in the future is decarbonised. It is unclear whether electricity generation capacity will be able to service both heating demand and a presumably still-increasing non-heating energy demand (Boardman, 2005, Department of Energy and Climate Change, 2009b).

4.3.3 Captive audiences and self-selection

One prevalent criticism of studies around sustainable lifestyles are ‘captive audiences’ involved in the collection of data (Keating, 1989). Interest in the nature of the people who take on energy-efficiency interventions is often separated from the measurements of energy use or installations, resulting in a skew of participants in the study conducted without stratification procedures in place (Crosbie, 2006, Guerin et al., 2000). The study of the energy reduction impact of interventions rarely measures the difference within participants; instead they are studying the differences between participants who can have many other differences in household characteristics outside of the physical and technical dimensions.

Detailed data can arise from experiments in housing involving demonstration projects in energy efficient methods of construction. Energy demonstration projects are not built as a mix of traditionally built and well-built, the self-selected occupant and the random occupant. Instead, the studies measure motivated occupants in well-built homes to a detail that can be rarely matched by studies involving homes that do not have any significant energy efficiency improvements. Other studies have measured the energy use of the improvements offered only to those homes that were part of the social housing stock. These measurements were subsequently used to verify building simulation models such as BREDEM in the 1980s and 90s (Shorrock et al., 1991, UK Energy Research Centre, 2008).

As stated earlier, the danger of using data from captive audiences occurs when researchers draw conclusions across the entire population of household using only this type of data. Self-selecting and motivated installers of energy efficiency measures are not a representative sample (Crosbie and

Baker, 2010). Neither are occupants of dwellings designed and advertised on the housing market as low-energy homes in demonstration projects that have extended end-use monitoring.

These dangers might well be rectified in the future with the exposure of a larger population to low-energy housing. Building regulations are to be tightened and the "Great British Refurb" for energy consumption is to be extended to a much larger segment of the population by ensuring that all homes are built to a zero-carbon standard by 2016 (Department for Communities and Local Government, 2008a). In a scenario of regulations requiring zero-carbon homes, the spread of the socioeconomic and lifestyles of occupants of low-energy homes will be extended to the entire population moving into new-build properties, removing the active choice of a low-energy home.

4.3.4 Secondary focus of investigation

Sometimes data on household consumption is collected as a secondary research question, notably in projects that are designed to measure the ability for on-site renewable generation to service the electricity needs of a home. This can lead to the output of, for example, a photovoltaic array, being measured, and the inputs of and the outputs to the electricity grid being measured, with the sum as the assumed electricity consumption of the household. However, the terminology is slightly different – instead of end-use energy as the measurement, the measurement of household (electricity) consumption corresponds to the energy balancing of the electricity grid in the presence of on-site generation (Firth et al., 2010).

4.3.5 Lack of ability to conduct longitudinal analysis

Studies that have large-scale breadth and are carefully planned unfortunately do not always ask the same questions relating to energy use to enable viable longitudinal analysis of changes to energy use over time. The most common split in survey questions is between asking for the metered energy consumption in housing surveys (U.S. Energy Information Administration, 2011, Department of the Environment Transport and the Regions, 2000b) and asking what the size of the last energy bill was in household surveys (Office for National Statistics and Department for Environment Food and Rural Affairs, 2010).

The other problem with the data is the infrequency of planned, stratified sampling of homes at a national scale. There has been a 15-year gap between surveys in the United Kingdom, and there is no guarantee that housing surveys will continue to cover energy consumption data as part of their design in the future according to the commissioning authority in England, the Department for Communities and Local Government (McIntyre, 2011). Surveys are also not designed to have cohorts available for analysis, as policy currently recommends keeping households on the survey panel for

two surveys and then rotating to a new set of households (Department of the Environment Transport and the Regions, 2002, Department for Communities and Local Government, 2010a).

The lack of data for longitudinal analysis also weakens both economic and psychological models of energy use. Economic models of energy use explain energy in terms of energy-relevant purchases and the ability of markets and policy to affect consumer behaviour. Psychological approaches to household consumption measures ways people are persuaded to use less energy (Lutzenhiser et al., 2010). Both model energy use as the sum of micro-level decisions made by people in households. Both focus on the individual decision-making process as either economic rationality or as a mental state, and lack understanding about externalities (European Central Bank, 2003) that influence choices in energy use.

4.3.6 Lack of identifying information

Validation of models of household energy use against actual energy use has historically been hampered by the requirement for privacy of residential occupants. Housing research in western democracies comes up against data protection laws that limit the amount of personal information available to the investigator. Data protection measures have been put in place for the built environment in the United Kingdom under the Data Protection Act 1998.

The Data Protection Act 1998 attempts to ensure that personal data, and especially sensitive[2] personal data is not misused in the United Kingdom. Academic research projects are subject to the regulations without any special exception or provision. There are eight principles that are laid down by the the Act. The first is that "any personal data must be processed lawfully and fairly." In an academic research context, the human subject of any investigation must give, and can take away at any time, his explicit permission to processing of personal data. There are exceptions to this requirement for permission, but the standards are high: protecting the vital interests of the data subject, the administration of justice, and the functions of a public nature exercised in the public interests by any person are examples of what would satisfy this requirement. Another principle

[2] Defined in the UK Data Protection Act 1998 as (a) the racial or ethnic origin of the data subject,
(b) his political opinions,
(c) his religious beliefs or other beliefs of a similar nature,
(d) whether he is a member of a trade union (within the meaning of the [1992 c. 52.] Trade Union and Labour Relations (Consolidation) Act 1992),
(e) his physical or mental health or condition,
(f) his sexual life,
(g) the commission or alleged commission by him of any offence, or
(h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings or the sentence of any court in such proceedings.
HMSO 1998. Data Protection Act 1998. *In:* Office, H. (ed.). London.

requires that personal data are “adequate, relevant, and not excessive in relation to the purpose or purposes for which they are processed.”

In the context of research in energy modelling of residential buildings, there are two main methods that would bring researchers into contact with personal data: surveys and meter readings. In the United Kingdom, the Building Research Establishment Domestic Energy Model that the current Building Regulations Part L was first based upon was built from the data obtained from participants who lived in purpose-built communities such as the Milton Keynes Energy Park in the 1980s, then on survey data in the 1996 English House Condition Survey (EHCS) (Henderson and Shorrock, 1986b, Energy Advisory Services, 1996, Anderson, 2002b, Sefton and Chesshire, 2005). However, the data in the EHCS does not include any spatial identifiers beyond the nation or English government office region. This has hampered sociological models of household energy use, as the patterns of consumption beyond the individual household level are more difficult to determine since the socioeconomic and built environment context of the household is difficult to investigate.

Surveys where participants voluntarily submit their meter data and any other data that may be required such as appliance ownership, type of heating system, or socioeconomic data on the occupants are widespread (Chapman, 1990, Environmental Change Institute, 1995, Chapman et al., 1985b, Alexander, 1983, Pears and Versluis, 1993, Zimmerman, 2009). However, they have been often limited to specific geographic areas, such as a housing estate, and have only recently become visible after pilot studies since the mid-1990s. The applicability of end-use monitoring has been limited to one sub-set of the residential sector, usually family owner-occupied housing that is more easily surveyed in the required detail (Lebot et al., 1995, Westergren et al., 1999).

One alternative approach that has been proposed by statisticians when high quality data inside of designed national-level surveys are restricted by data protection is the concept of area classification. Between 2002 and 2007, a general purpose area classification was created for the Office of National Statistics by the University of Sheffield (Vickers and Rees, 2007). The classification groups geographically-defined areas according to characteristics that are held to be common to the population in each group. These groupings are also named clusters. They were generated from 2001 census data and have been accepted as a format for national statistics for the United Kingdom (Office for National Statistics, 2005a). This alternative way of partial identification of households is to be explored extensively in this thesis, since data matching to area classifications, but not individual census areas, of housing surveys in England has been made possible by the Department for Communities and Local Government.

4.3.7 Data available in aggregate

In the energy sector, the highest quality of measurement of energy use is contained in aggregate data, or data collected by group defined either by geographical area or by electricity or gas substation. These data are used to build group-level statistics, such as the total amount of energy used, or the percent of households that have two people, from individual-level variables. It is pulled from all the aggregate units and is not a sample, and is based on the measurements for all individuals with the group. Using means of groups to predict the behaviour of individuals commits an ecological fallacy (Robinson, 1950, Freedman, 1999). In situations where individual data is available in anonymised form with some group information, then individual households can be clustered into area classifications for comparisons using individual and group characteristics. Correlations between households in the same classifications are more likely for variables selected for area classifications, and likely reducing the variance between the mean of these variable and individual values belonging to the same area classification. Area classifications and clustering were introduced in Chapter 3.

Aggregate data is created by the totalling of all individuals into discrete groups, most often geographic groups, in order to protect the identity of the individuals. This is often due to the data being taken without the express permission of the participants, as is the case with energy use statistics (Elexon, 2010), or is compulsory, as is in the case with national censuses such as the one that is run in the United Kingdom.

There has also been a recent growth in applied mathematical modelling of the relative energy use of urban areas. Recent research that used a proxy of miles of electrical cabling to represent aggregate energy consumption concluded that there is a growth model that can predict energy consumption using the size of the urban area as a predictor (Bettencourt et al., 2007). This did not take into account the amount of natural gas or non-electric fuels used for heating in different urban areas and therefore the confounders on non-heating energy end-use were not considered. However, this size scaling model has been disputed, and an urban hierarchy model was proposed, claiming that it better fit the data (Shalizi, 2011). This research is advancing the understanding of the relationship of the relative size of urban areas and resource consumption in general, but there is limited predictive power available for the individual household because the level of aggregation is the urban area.

4.3.8 Conclusions and recommended criteria for the handling of data for a project examining individual and area level effects on non-heating end-use energy

After a review of the criticisms of data produced to support the modelling of single dwellings or households, criteria were established for the selection of data in order to investigate a combination of individual and area level effects on non-heating end-use energy.

Recommendation 1 – Individual level demographic data. The data must be able to interact with the main physical-technical-economic model of household energy use. This will require an interface with the BREDEM model of a single dwelling and must be able to address the concerns of the writers of building regulations in assuming a household's socioeconomic characteristics. The data will include:

- Physical and human measures of house size
- Demographics of the household that affect non-heating end-uses

Recommendation 2 – Individual level - Recorded metered consumption data. The data should record actual energy consumption, not the count of installations feeding into a figure for modelled energy consumption. This will include:

- Annual meter –recorded energy consumption by end-use and/or fuel
- Fuel for heating end-uses

Recommendation 3 – Individual and area level - Sampling methods. The data must be designed to cover the entire population of a large geographic area such as a region or nation and also designed to randomly select participants instead of selecting on the basis of availability. This will include:

- Sampling methods stated, preferably stratified sampling if done at a nationwide scale
- Total number of homes included

Recommendation 4 - Individual and area level - Direct measurement. End-use energy should be directly measured in the data, and not be an estimated or derived measure. This will include:

- Direct measurement of end-use energy

Recommendation 5 – Individual and area level - Longitudinal data. The dataset should have a set of questions about households that are stable over time, include recorded end-use energy, and collected at regular intervals. This will include:

- Longitudinal elements

Recommendation 6 – Individual level - Case identification. The data should include sufficient information to fully identify the case within a local geographic area, or if data protection measures prevent this, partially identify the case within an area classification. This will include:

- Address file
- Area classification

Recommendation 7 – Area level – Aggregate data. The data should be able to be verified using aggregate data that covers the non-heating end-use energy of the entire population with extensive detail on the characteristics of the population. This will include:

- Aggregated physical measures of dwellings and human measures of household size
- Aggregated demographics of households that affect non-heating end-uses
- Total annual meter –recorded energy consumption by end-use and/or fuel
- Aggregated fuel for heating end-uses
- Total number of homes included
- Area classification

4.4 Sources of Data

4.4.1 Introduction

This section will introduce the data that is available to the research project and use the recommendations developed earlier to select the best data available. It is anticipated that no one set of data available will fulfil all the criteria outlined above as the nature of the project involves individual household level data and area level data, and the exact area of an individual household will be masked by data protection law. The reasons for selecting datasets that cover the UK have been introduced earlier. These reasons are that all segments of the residential population of the UK are available for study in both aggregated and disaggregated forms, and the relatively high correlation found in the UK in comparison with other developed countries between delivered electricity consumption and non-heating end-use energy.

The data is broadly grouped into three groups. The first are surveys that are specialised in measuring building performance. The second are surveys that are broader, non-specific housing surveys. The third are surveys that use aggregate data for energy use in households. After a brief introduction to each dataset, the strengths and weaknesses of the datasets will be summarised and datasets will brought forward for investigation in this thesis.

4.4.2 Specialist housing surveys in energy use – individual household level

There are many specialist surveys that have been made in the context of the United Kingdom. Specialist building performance surveys have been organised for the sole purpose of measuring the performance of buildings. A building simulation in the UK is a heat balance equation. The majority of effort and time in measuring building performance has two outcome variables: energy use and

indoor temperature (Hong et al., 2006a). Most specialist surveys are made with an academic or an engineering focus, and are exclusively designed to support the physical-technical-economic model of energy use. They have been used as specialist evidence in the past to support building regulations that govern the energy performance of buildings, but not specifically to predict non-heating end-uses (Shorrock et al., 1991, Department for Communities and Local Government, 2008b, Sefton and Chesshire, 2005). The surveys that were identified for investigation were:

- BREDEM version 1996 Validation Data
- Pennyland Project
- York Energy Demonstration Project
- Monitored Domestic Energy Use Data Archive
- Hull Low Energy Housing Project
- Carbon Reduction in Buildings (CaRB) Home Energy Survey
- Homes Energy Efficiency Database
- Warm Front Database

4.4.3 Non-specific housing surveys – individual household level

Non-specific housing surveys are designed to sample a wide variety of housing conditions to develop and inform departmental housing policies. They have been developed by government departments in conjuction with outside agencies, notably the Office for National Statistics and the Building Research Establishment to identify questions that can integrate with surveys that investigate more specific areas for research (Department for Communities and Local Government, 2011b).

The surveys that were identified for investigation were:

- English Housing Survey
- Survey of English Housing
- English House Condition Survey
- Living Costs and Food Survey

4.4.4 Non-specific housing surveys – area level

Aggregate data on energy use and households are collections of data that include the entire population of households and are not a sample of households. In order to protect the privacy of respondents, all data is summarised by geographic area. There are also classification systems that estimate the characteristics of clusters of either continuous or non-continuous groupings of areas and market segmentation that estimates the number of households belonging to each grouping per geographic area. The datasets that were investigated were:

- 2001 United Kingdom Census
- Department for Energy and Climate Change Small Area Statistics
- Office for National Statistics Area Classifications
- Energy Saving Trust Market Segmentation
- Experian Mosaic Market Segmentation

4.5 Assessment of data sources

Each of the data sources were assessed according to the criteria established earlier in this section. The following headings will be investigated:

Table 4.2: Investigation of data sources

Basic information and availability	Criteria for selection
Description Dates of collection Number of respondents Open access to raw data	Demographic data (Individual) Recorded metered consumption data (Individual) Sampling methods (Individual / area) Direct measurement (Individual / area) Longitudinal data (Individual / area) Case identification (Individual) Aggregate data (Area)

4.5.1 Basic Information and Availability

Table 4.3: Specialist housing surveys of energy use

Dataset and Date	Description	Respondents	Open access to raw data
Specialist housing surveys of energy use – individual household level **BREDEM 1996 Validation Data**			
BRE (date unknown)	Standard BRE testbed homes : 4 passive solar homes at Linford and 10 other homes	14	No, but available in summary in (Shorrock et al., 1991)
Washington, Co Durham (date unknown)	New low energy, well insulated terraced homes	6	No, but available in summary in (Shorrock et al., 1991)

Dataset and Date	Description	Respondents	Open access to raw data
Birmingham Energy Improvement Kit (1979-82)	Birmingham City Council housing project to install a new kit in existing dwellings with large amounts of social data	25	Yes, see Monitored Domestic Energy Use Data Archive (1973-83)
Sandwell, Birmingham (date unknown)	Energy efficiency demonstration project in a high rise block heated by electric storage	Unknown	No, but available in summary in (Shorrock et al., 1991)
Collyhurst, Manchester (date unknown)	Demonstration project of homes with gas fired condensing boilers	30	No, but available in summary in (Shorrock et al., 1991)
Milton Keynes Energy Park (1990)	Designated in 1985, the Energy Park was planned as an international demonstration project of energy efficiency. All buildings constructed in the Energy Park were required to demonstrate high levels of energy efficiency.	25	Partially available due to degradation of data in storage (Summerfield et al., 2007)
Super Insulated, Milton Keynes (date unknown)	4 super-insulated and 4 other homes at Two Mile Ash in Milton Keynes. This was a demonstration project funded by the European Commission and the Polytechnic of Central London (now Westminster University)	8	No, but available in summary in (Shorrock et al., 1991)
Specialist housing surveys of energy use – individual household level Additional			
Pennyland , Milton Keynes (1976)	Monitoring of an estate of low energy houses in Milton Keynes to study the possibility to produce a mass-market low energy house. Created by the Open University Research Group.	177	No
York Energy Demonstration Project (1991-94)	Project that studied the technology available to local authority-owned housing stock in York City Council	230	No

Dataset and Date	Description	Respondents	Open access to raw data
Monitored Domestic Energy Use Data Archive (1973-83)	Collection of studies conducted by universities across the United Kingdom, including the Better Insulated Housing Programme and other one-off studies in building insulation. Social housing only.	unknown	Yes (Building Research Energy Conservation Support Unit, 2007)
Hull Low Energy Housing Project (1981)	Collection of data on energy consumption and internal dimensions of social housing in Hull City Council	150	Yes (Pearson, 1981)
Carbon Reduction in Buildings (CaRB) Home Energy Survey (2007)	Research project that aimed to identify the socio-technical causes of domestic energy consumption	427	Yes (Shipworth, 2010)
Homes Energy Efficiency Database (ongoing)	The Homes Energy Efficiency Database (HEED) is a "survey of surveys" that recorded the number of installations related to energy efficiency in the housing stock	10 million	Yes, but not location data.
Warm Front Database (2001-03)	Database of energy consumption and indoor temperature of households that had applied for a Warm Front grant in five urban areas	1,500	Yes, but not energy data

Table 4.4: Non-specific housing surveys

Dataset and Date	Description	Respondents	Open access to raw data
Non-specific housing surveys – individual household level			
English Housing Survey (2009-)	The EHS is a national survey commissioned by the Department for Communities and Local Government (DCLG) that collects information about current housing circumstances and the condition and energy efficiency of housing in England.	17,000	Yes
English Housing Survey Energy Follow-up Study	A sub-sample of the English Housing Survey will collected extensive	Unknown	Yes; beyond timescale of this

Dataset and Date	Description	Respondents	Open access to raw data
(2011/12)	information about the energy habits and consumption of households.		research
Survey of English Housing (1993-2008)	The SEH was a multi-purpose housing survey which provided a comprehensive range of basic information on households and their housing	19,000 (2008)	Yes
English House Condition Survey (1967-2008)	The EHCS was a physical housing survey that was conducted by surveyors focussing on physical factors and market value	25,000 per annum (1996)	Yes
English House Condition Survey Fuel Survey Sub-sample (1986-2001)	These sub-samples all collect actual meter readings from either the property itself, or from the energy company supplying the property.	2,531 (1991)	Yes to 1996 (Department of the Environment Transport and the Regions, 2000b)
Living Costs and Fuel Survey (2008)	As part of the new framework of the Integrated Household Survey, the LCF is a continuous survey of household expenditure on many uses including energy bills, food consumption and income.	6,140	Yes (Office for National Statistics and Department for Environment Food and Rural Affairs, 2010)
Non-specific housing surveys – area level			
United Kingdom Census (2001)	The Census is a count of all the population and households in the country, with additional data about the household and their occupants. The 2011 Census will be made available at the level of Super Output Areas in 2013-14.	22,539,000 households in England	Yes (aggregated)
Department for Energy and Climate Change Small Area Statistics (2005-2009)	The DECC small area statistics database includes the total amount of energy consumed by the residential sector, collected from electricity and natural gas companies.	22,886,265 electricity meters (2009)	Yes (aggregated)

Dataset and Date	Description	Respondents	Open access to raw data
Office for National Statistics Area Classifications (2007)	The ONS area classifications area a non-specific system of classifying output areas.	175,434 output areas	Yes
Energy Saving Trust Market Segmentation (2007)	The Energy Saving Trust Market Segmentation is a classification system of the propensity and financial ability of people to install energy efficiency measures in their homes (Energy Saving Trust, 2009a).	1,602,314 postcodes	No
Experian Mosaic Market Segmentation (2009)	Mosaic is a general-purpose market segmentation system that estimates the demographics, behaviour, and lifestyles – there is a tier system that can be used at the individual, household or postcode level.	N/A	Yes, at household level

After a survey of availability of the data, the possible datasets were narrowed to the following shortlist:

Table 4.5: Shortlist of datasets

Type	Dataset
Specialist housing surveys in energy use – individual household level	Monitored Domestic Energy Use Data Archive (1973-83) Hull Low Energy Housing Project (1981) Carbon Reduction in Buildings (CaRB) Home Energy Survey (2007) Homes Energy Efficiency Database (ongoing)
Non-specific housing surveys – individual household level	English Housing Survey (2009-) Survey of English Housing (1993-2008) English House Condition Survey (1967-2008) English House Condition Survey Fuel Survey Sub-sample (1986-2001) Living Costs and Fuel Survey (2008)
Non-specific housing surveys – area level	United Kingdom Census (2001) Department for Energy and Climate Change Small Area Statistics (2005-2009) Office for National Statistics Area Classifications (2007) Experian Mosaic Market Segmentation (2009)

4.5.2 Assessment of shortlist of datasets against selection criteria

This section will assess the shortlisted data against the recommended criteria established earlier in the thesis. Each criteria will be qualitatively scored using a modified Likert scale (Likert, 1932):

- Strongly meets the criterion **(++)**
- Somewhat meets the criterion **(+)**
- Neither meets nor breaks the criterion **(0)**
- Somewhat breaks the criterion **(-)**
- Strongly breaks the criterion **(--)**

There is no overall score for each dataset as the scores are not assumed to represent an interval measurement of the suitability of the dataset.

Table 4.6: Qualitative quality assessment of the shortlist of datasets

Monitored Domestic Energy Use Data Archive (1973-83) (Building Research Energy Conservation Support Unit, 2007)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	"The data were physical rather than sociological. An attempt has been made in the printed project appendices to present supplementary information which might be needed by a user to interpret the variables held on computer files. In this context many will be particularly interested in the supplementary information which can be generally described as sociological, - details of occupancy, social status and income of householder etc."	**+**
Recorded metered consumption data (Individual)	Energy consumption was recorded, with the aim of assessing the affects of thermal insulation in houses.	**+**
Sampling methods (Individual / area)	The projects were selected ad hoc, and were only drawn from housing under the control of the public sector	**--**
Direct measurement (Individual / area)	In all cases, energy consumption was measured and not estimated.	**++**
Longitudinal data (Individual / area)	The studies were not repeated yearly, but many of the studies did take assessments before and after the implementation of measures to reduce heating, but not non-heating, end-uses , notably the installation of insulation.	**0**
Case identification (Individual)	Some of the projects are limited to a single geographic area, making the link to a geographic area implicit. Others were of social housing generally inside of a local authority. Some inferences on the type of areas surveyed can be made.	**0**
Aggregate data (Area)	Not aggregate data.	**0**

Hull Low Energy Housing Project (1981) (Pearson, 1981)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	"The social survey was undertaken to provide information on behavioural aspects of domestic energy use in council housing. The survey is part of the Low Energy Housing Project which aims to combine the building of low energy housing with monitoring energy use and investigating the determinants of domestic energy consumption."	**++**
Recorded metered consumption data (Individual)	Fuel consumption and expenditure data were collected. However, the recording of data took place only between April and June of 1981 so no annualised energy estimates could be made.	**-**
Sampling methods (Individual / area)	The project only drew from housing under the control of the public sector.	**--**
Direct measurement (Individual / area)	In all cases, energy consumption was measured and not estimated.	**++**
Longitudinal data (Individual / area)	This study was not repeated.	**--**
Case identification (Individual)	The project was limited to a social housing inside of a single local authority, making an implicit link to both the type of area in which the house was located and the a geographic area.	**0**
Aggregate data (Area)	Not aggregate data.	**0**

Carbon Reduction in Buildings (CaRB) Home Energy Survey (2007)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	Research project carried out by several institutions that was aimed, among many things, at identifying the socio-technical causes of domestic energy consumption. This included socio-economic and building energy efficiency variables.	**++**
Recorded metered consumption data (Individual)	Meter readings were recorded on site by the interviewer and then some participants recorded meter readings once a quarter for the next year, and some historical meter readings were provided by the UK Department for Business Innovation and Skills in the last two years.	**++**
Sampling methods (Individual / area)	Stratified (by Government Office Region and socio-economic status) random sample. 54 postcode sectors were chosen, and 21 addresses sampled in each postcode sector.	**++**

Carbon Reduction in Buildings (CaRB) Home Energy Survey (2007)		
Direct measurement (Individual / area)	Recorded meter data was used in this survey.	**++**
Longitudinal data (Individual / area)	Replicated some variables from the 1981-85 Human Factors Study in Domestic Gas Consumption (heating end-uses only) and DEFRA attitude questions, which added longitudinal elements to the survey. However, these elements are focused on heating end-uses.	**0**
Case identification (Individual)	Each case can be identified to the postcode sector, but data protection rules may preclude re-use of the data for an additional purpose, including having access to the address file of each respondent	**+**
Aggregate data (Area)	Not aggregate data.	**0**

Homes Energy Efficiency Database (Energy Saving Trust, 2008, Bruhns et al., 2011)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	HEED collects surveys of energy efficiency and on-site electricity generation installations, for example, cavity wall insulation and solar hot water, along with surveys of the property, but not the occupants, concerned. Focus on heating end-uses.	**0**
Recorded metered consumption data (Individual)	Electricity and gas bills, but not meter readings, for 13 million dwellings.	**+**
Sampling methods (Individual / area)	The HEED database does not have a sampling method and does not include homes that do not have installations.	**--**
Direct measurement (Individual / area)	The HEED database records the improvement in designed energy consumption directly from energy bills.	**++**
Longitudinal data (Individual / area)	The database depends on the questions asked in individual schemes and surveys. No standardisation of the surveys.	**--**
Case identification (Individual)	The location of individual homes is only available to contractors and is not available to the research community.	**--**
Aggregate data (Area)	Not aggregate data.	**0**

English Housing Survey (Department for Communities and Local Government, 2011b)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	The English Housing Survey contains a wealth of both demographic information on the household and the dimensions and structural information on the dwelling	**++**
Recorded metered consumption data (Individual)	No recorded consumption data.	**--**
Sampling methods (Individual / area)	Stratified by housing tenure (e.g owner-occupied, socially rented)	**+**
Direct measurement (Individual / area)	No recorded consumption data will be available until 2012/13.	**--**
Longitudinal data (Individual / area)	The survey is designed to include questions that were previously part of the Survey of English Housing and the English House Condition Survey. Households that take part in the survey are not retained for subsequent years of the EHS, but have been asked to participate in an energy follow-up survey. No decision has been made on including the follow-up survey in future years.	**+**
Case identification (Individual)	Individual homes may not be identified. Data matching may be available to place homes into area classifications, but not into individual census areas such as Super Output Areas.	**+**
Aggregate data (Area)	Not aggregate data.	**0**

Survey of English Housing (Department for Communities and Local Government, 2011b)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	The Survey of English Housing contains a wealth of demographic data on each household in its database, but no data on the internal dimensions of the dwelling.	**+**
Recorded metered consumption data (Individual)	No recorded consumption data.	**--**

Survey of English Housing (Department for Communities and Local Government, 2011b)		
Sampling methods (Individual / area)	Stratified by government office region, housing tenure (e.g owner-occupied, socially rented), and socioeconomic class, specifically if they are or are not in the higher social classes	**++**
Direct measurement (Individual / area)	No recorded consumption data.	**--**
Longitudinal data (Individual / area)	The survey does not have any built-in longitudinal features. However, it does ask a standard set of questions to each year's group of household surveys.	**-**
Case identification (Individual)	Individual homes may not be identified. Data matching may be available to place homes into area classifications, but not into individual census areas such as Super Output Areas.	**+**
Aggregate data (Area)	Not aggregate data.	**0**

English House Condition Survey (Department for Communities and Local Government, 2010a)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	The English House Condition survey contains a wealth of data on the dimensions of the dwelling, and additionally has extensive information on the demographics of the household.	**++**
Recorded metered consumption data (Individual)	A fuel sample was collected along with the survey until 2001. This collected nine quarters of metered energy usage from a sub-sample of the homes included in the overall survey. (The last survey that is available to all academic researchers was in 1996.)	**++**
Sampling methods (Individual / area)	Stratified by government office region, housing tenure (e.g owner-occupied, socially rented), building age, and building type (e.g. converted flat, detached house)	**++**
Direct measurement (Individual / area)	Participants in the survey were trained to read and record energy meters.	**++**
Longitudinal data (Individual / area)	The survey does not have any built-in longitudinal features. However, it does ask a standard set of questions to each year's group of household surveys and retains a dwelling in the survey for two years.	**0**
Case identification (Individual)	Individual homes may not be identified. Data matching may be available to place homes into area classifications, but not into individual census areas such as Super Output Areas.	**+**

English House Condition Survey (Department for Communities and Local Government, 2010a)		
Aggregate data (Area)	Not aggregate data.	**0**

Living Costs and Fuel Survey (Office for National Statistics and Department for Environment Food and Rural Affairs, 2010)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	The survey contains information about the demographics of the household, but not the internal dimensions of the dwelling in which they live.	**+**
Recorded metered consumption data (Individual)	No recorded meter data, but the latest bill, billing cycle, and payment type are included to approximate energy usage, and this data is available up to 2009 (The latest fuel survey of the English House Condition Survey was collected in 2001, and the last survey that is available to all academic researchers was in 1996.)	**-**
Sampling methods (Individual / area)	The LCFS has a stratified sample based on government office region and socioeconomic group.	**++**
Direct measurement (Individual / area)	No recorded meter data.	**--**
Longitudinal data (Individual / area)	The survey does not have any built-in longitudinal features. However, it does ask a standard set of questions to each year's group of household surveys.	**+**
Case identification (Individual)	Individual homes are not identified, but the area classification at output area level is released as part of the dataset.	**+**
Aggregate data (Area)	Not aggregate data.	**0**

United Kingdom Census (ESRC Census Programme, 2006)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	Not an individual dataset. However, all of the demographic data for the aggregate population is present, including extensive demographic information on households and some information on the internal dimension of dwellings.	**+**
Recorded metered consumption data (Individual)	Not an individual dataset.	**0**
Sampling methods (Individual / area)	The entire population of the United Kingdom is required to complete the census questionnaire.	**++**
Direct measurement (Individual / area)	Energy use is not part of the census questionnaire.	**--**
Longitudinal data (Individual / area)	The design of the census every ten years retains questions, especially on socioeconomic background, but also on the availability of central heating.	**++**
Case identification (Individual)	Not an individual dataset. Individual entries in the census are not available until 100 years afterwards.	**0**
Aggregate data (Area)	All homes are included, and the statistics are aggregated into small areas – output areas (125 households), lower layer super output areas (1500), and middle layer super output areas (7200)	**++**

Department for Energy and Climate Change Small Area Statistics (Department for Energy and Climate Change, 2010b)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	Not an individual dataset.	**0**
Recorded metered consumption data (Individual)	Not an individual dataset.	**0**
Sampling methods (Individual / area)	The entire population of natural gas and electricity meters are included in this database. The companies that are obliged to manage the balancing and settlement arrangements for the natural gas and electricity networks provide the information on the consumption of every meter in the country.	**++**

Department for Energy and Climate Change Small Area Statistics (Department for Energy and Climate Change, 2010b)		
Direct measurement (Individual / area)	Every energy meter (gas, ordinary electricity, and economy7) is included in the database. The database measures consumption over the course of a financial year (6 April – 5 April). If the billing and recording cycle does not match the financial year, some consumption is estimated.	**+**
Longitudinal data (Individual / area)	Aggregated consumption figures have been released for the same census areas every year since 2005 and are now considered an official national statistic.	**++**
Case identification (Individual)	Not an individual dataset.	**0**
Aggregate data (Area)	All homes are included, and the statistics are aggregated into small areas - lower layer super output areas (1500 households), and middle layer super output areas (7200)	**++**

Office for National Statistics Area Classifications (Office for National Statistics, 2008b)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	Not an individual dataset.	**0**
Recorded metered consumption data (Individual)	Not an individual dataset.	**0**
Sampling methods (Individual / area)	The area classification system was created by the Office of National Statistics after the 2001 Census. The classification uses cluster analysis to reduce 41 socio-economic and built environment census variables to a single indicator. (Vickers and Rees, 2007)	**++**
Direct measurement (Individual / area)	Energy use was not considered in the creation of this classification system.	**--**
Longitudinal data (Individual / area)	This system of classification can be repeated after every census to take account of the changes in the socio-economic circumstances of a census area.	**++**
Case identification (Individual)	Not an individual dataset.	**0**

Office for National Statistics Area Classifications (Office for National Statistics, 2008b)		
Aggregate data (Area)	Area classifications were created for every census area in the United Kingdom – output areas (125 households), lower layer super output areas (1500), and middle layer super output areas (7200)	**++**

Experian Mosaic Market Segmentation (Experian UK, 2009)		
Selection criteria	**Qualitative Assessment**	**Score**
Demographic data (Individual)	Not an individual dataset.	**0**
Recorded metered consumption data (Individual)	Not an individual dataset.	**0**
Sampling methods (Individual / area)	54% percent of the data used to construct Mosaic is sourced from the 2001 Census and the other 46% comes from sources such as the Experian Lifestyle Survey, consumer credit databases, the electoral roll, shareholder registers, Land Registry data, Council Tax information, the British Crime Survey, Expenditure and Food Survey, the Health Survey for England, and other sources. The exact methods used to construct the classification system, however, is commercial property and not subject to academic scrutiny.	**+**
Direct measurement (Individual / area)	Energy use was not considered in the creation of this classification system.	**--**
Longitudinal data (Individual / area)	This system of classification can be repeated to include new data yearly, especially from tax rolls, lifestyle surveys, and credit databases, to take account of the changes in the socio-economic circumstances of a census area.	**++**
Case identification (Individual)	Not an individual dataset.	**0**
Aggregate data (Area)	The Mosaic database predicts the likely socioeconomic classification of all the households in a postcode, then produces counts of likely households of each classification by lower layer super output area (1500 households), and middle layer super output areas (7200)	**++**

4.5.3 Selection of datasets to use in this investigation

The selection of datasets for this investigation was guided by the assessments of each shortlisted dataset above. Some datasets had 'fatal flaws' from self-selection or the lack of inclusion of energy consumption as a measured, or even estimated, variable in the dataset. The choice of aggregated datasets was straightforward with the inclusion of the entire population of England in the 2001 Census and the DECC small area consumption statistics database. The choice of classification was contested between established, well-funded, but commercially sensitive classification systems and newer systems funded by the public sector and open to scrutiny. Lastly, the decision on the individual level dataset was closely contested between a newer, well-designed dataset developed in academia and an older, also well-designed, but much larger dataset developed by central government in England. The results of this selection process is an individual level dataset, an area level dataset, and an area classification system for investigating the possibility of including both individual and area level predictors of non-heating end-use energy.

Many individual level datasets that focus on the recording of metered energy use are flawed by limitations on the location and type of homes that they can assess. The Monitored Energy Use Data Archive and the Hull Low Energy Project only had access to a single estate, a single local authority, or to publicly-owned housing. The final result of this research work is to propose a model at the national, and not the local scale, and to encompass all housing, and not just social tenures. These datasets were not considered any further. The Homes Energy Efficiency Database only included homes that had energy efficiency or generation installations, and therefore the database is solely a self-selected sample of households, housing associations, or local authorities that are concerned enough about energy efficiency to commit funds to install these measures. Although Bruhns, Hamilton et al. (2011) found that these facts did not too greatly skew the distribution of households, even aggregated location data was only available to government contractors, so this database was not considered further.

Cases with electric heating in housing survey datasets need to be minimised. For such houses it is impossible to separate out end-uses in their reported electricity consumption. Similarly, small statistical areas that report a significant proportion of dwellings with no central heating will be assumed to have higher than acceptable electricity use for heating to be considered in validation and refinement procedures in the future. Typical examples of areas that have high amounts of electric heating installed are tall buildings and areas that have shifted from commercial to residential use (Hassell and Olivier, 2005, Pank et al., 2002). Therefore, high-density urban areas are likely to be underrepresented in these populations.

Aggregated socioeconomic demographic information, internal dimensions of buildings, and energy use of electricity in the residential sector are all available to researchers and these datasets will be brought forward in this thesis. The 2001 United Kingdom Census is a mandatory questionnaire sent to the entire population of the country and conducted in March 2001. The census includes for each census area a detailed and near-complete record of households (Simpson and Brown, 2008). This record includes details of the size of the dwelling, the heating arrangements of the dwelling, the economic activity of the household, and the age range of both the dwelling and the members of the household, amongst many other variables measured in the census. The census does not distinguish between the household and the dwelling. The Department for Energy and Climate Change Small Area Statistics database includes the total energy consumption of a census area by meter type collected directly from the balancing and settlement arrangements for the wholesale and retailing of energy to the residential sector (Elexon, 2010). The types of meters available are natural gas, ordinary electricity, and economy7 electricity meters. This data is only available for the residential sector, as commercial or industrial users, who are fewer in number and higher in energy consumption per address, could be individually identified if census area-level detail was released. The combination of aggregated census and total domestic consumption data provides a powerful database on the total consumption of the population with much less concern about self-selection and non-response issues.

The selection of an area classification system was a close decision between the official national statistic promoted by the Office for National Statistics and the well-established commercial market segmentation promoted by the specialist data company Experian. Although the dataset assembled by Experian for its Mosaic segmentation dataset is extensive and combines many other datasets together, it suffers from the lack of available documentation on the methodology for determining the classifications and the assignment of postcodes. The area classification system created by the Office for National Statistics with the University of Sheffield is open to scrutiny, describes in detail the variables used and the cluster analysis methodologies employed, along with assumptions, limitations, and suggestions for researchers to both use and improve on the work. As it provides for a national statistic, it is possible to request data matching of individual datasets to ONS area classifications of individual census areas if the census area is not released due to data protection rules. As the census area data of the Experian Mosaic dataset is a count of households, data matching would entail creating a second layer of methodology for classifying each area based on the count estimates created out of blocks of postcodes. Therefore, the ONS area classification system was selected for this project.

The choice of the individual household level dataset was also a close-run decision between the 2007 Carbon Reduction in Buildings Home Energy Survey (CaRB survey) and the fuel sub-sample of the 1996 English House Condition Survey (EHCS). The CaRB survey was a well-designed, stratified random sample, but it was a survey specifically designed to assess the impact of indoor temperature on heating end-use energy consumption (Shipworth, 2010). The EHCS was also a stratified sample, using four categories for stratification, crucially using tenure as one of them in line with all future housing surveys and energy follow-up surveys, in contrast to two categories for stratification in the CaRB survey which did not include physical characteristics of dwellings. In addition to the focus on heating end-uses in the CaRB survey, the CaRB survey was much smaller (427 as opposed to almost 2,531 in the EHCS), therefore the likelihood of finding statistically significant results is smaller because of the reduced sample size. Access to the CaRB survey is also somewhat restricted due to the agreement between the participants and the original researchers on the re-use of data for purposes outside of investigating heating end-use energy. Therefore, the EHCS fuel sub-sample was selected, as it is an open dataset, and a process of data matching of ONS area classifications that protected the anonymity of the participants to each case in the fuel sample of the 1996 English Housing Condition Survey was conducted for this project by the Department for Communities and Local Government (McIntyre, 2011) .

There were two problems with the EHCS database that are addressed later in this thesis. The first was the age of the dataset and the need to compare it with more recent aggregated datasets collected 5-10 years later. This thesis will propose to use energy billing data from the past decade, notably from the 2008 Living Costs and Fuel Survey to estimate the energy consumption of the participants in the 1996 EHCS and to compare to the aggregated energy use totals found in the DECC Small Area Statistics Database. The second was the lack of geographic location of the individual cases. This thesis matches the data to area classifications and government office regions to simulate the effect of location, using the built environment criteria for setting area classifications as described in the methodology released by the ONS and the University of Sheffield.

The conclusion of this analysis is that the following datasets were taken forward for further investigation in the project:

Individual Household Level

- 1996 English House Condition Survey, Fuel sub-sample
- 2008 Living Costs and Fuel Survey

Area Level

- 2001 Census
- 2008 Department of Energy and Climate Change Small Area Statistics Database

Classification

- Office for National Statistics 2001 Area Classification

Chapter 5 - Methodological approaches

5.1 Introduction

This section will outline alternative statistical methodologies that are available for assessing the non-heating end-use energy of households for detailed comparison in Chapter 6.

This thesis aims to develop:

- a method for updating of models of domestic energy use between major surveys, and
- a method for using area-based variables in models of domestic energy use.

In Chapter 4, a range of data needed for analysis was selected using both individual household level and area level variables. A number of recommendations were made for the handling of such data (section 4.3.8). In this chapter, these recommendations feed into the range of available methods. These methods will be built around two timescale options: one that can be updated annually, and the other that is updated decennially (once every decade).

In Chapter 3, statistical methods for modelling energy use in buildings used in the UK over the last 30 years were introduced. There is a range of statistical methodologies for the modelling of the energy use of an individual household that were considered in addition to the established methods from the macroeconomic and parametric "families" of statistical analysis that could allow area-based variables.

The conclusions of this chapter will recommend a range of statistical models for detailed comparative analysis in Chapter 7.

5.2 Review of data recommendations

Chapter 4 outlined general recommendations for evidence that would best serve the stated aims of this thesis. These were:

- **Recommendation 1 – Individual level - Demographic data**
- **Recommendation 2 – Individual level - Recorded metered consumption data**
- **Recommendation 3 – Individual and area level - Sampling methods**
- **Recommendation 4 – Individual and area level - Direct measurement**
- **Recommendation 5 – Individual and area level - Longitudinal data**
- **Recommendation 6 – Individual level - Case identification**
- **Recommendation 7 – Area level - Aggregate data**

The table below reviews the purpose for including each type of data in models of domestic energy consumption.

Table 5.1: Recommendations for data to be included in modelling of non-heating end-use energy in dwellings

Recommendation for data	Purpose
Demographic data	• To measure the physical size of the household • To measure the impact of income on energy use • To measure the impact of the number of occupants • To measure the physical size of the household
Sampling methods	• To determine if non-heating end-use energy equates to electricity use in the household • To measure the impact of housing typology
Direct measurement / recorded metered consumption data	• To measure directly the non-heating end-use energy of a dwelling
Longitudinal data	• To measure the seasonal and annual variation of energy use
Case identification	• To have a measure of intensity of the urban area • To measure the regional dimension of energy usage
Aggregate data	• To obtain a general classification of household socioeconomic status • To obtain a general classification of household socioeconomic status

5.3 Future updating

There are two options that emerge from the analysis of the data for future updating – a model that is designed to be updated yearly, but is more experimental and subject to error, and a model that is designed to be more stable and updated once every decade. The two different options are driven by the data available in late 2010 and before data releases from the 2011 Census at the level of LLSOAs, and the 2011 English Housing Survey and its follow-up energy survey anticipated in 2013-14.

Detailed surveys of energy use in dwellings are collected infrequently. Previous models have been left unchanged for over a decade because of the sporadic nature of government commissioning of surveys. Presently, one large-scale, government-run survey of domestic energy use has been made since 1996, and another survey is underway. Neither dataset is yet available for academic research (BRE, 2005a, Department for Communities and Local Government, 2011b).

Datasets were selected in Chapter 4 out of the recommendations made for the availability and applicability of data that would enable the development of a model with both area-based and individual household-based variables. These were:

- Fuel sub-sample of the 1996 English House Condition Survey
- 2008 Living Costs and Fuel Survey
- 2001 Census
- 2008 Department of Energy and Climate Change Small Area Statistics Database
- Office for National Statistics 2001 Area Classifications

5.3.1 Option 1: Annual model options

One option for the models is to update them based on information on domestic electricity consumption collected yearly. This model would be adaptable in future years as a yearly update, as the 2011 Census and future energy follow-up surveys of the English Housing Survey (EHS) will be derived from the same core set of question as the Living Costs and Fuel Survey (LCFS) as part of the Integrated Housing Survey introduced in 2008. Living Costs and Fuel Survey asks for the value of the last fuel bill but does not directly measure electricity consumption. The LCFS also uses the European Standard Classification of Individual Consumption by Purpose that establishes common terms of reference for collecting data on expenditure and other consumption throughout the European Union, making it easier to adapt this model to other settings within Europe.

5.3.2 Option 2: Decennial model options

A second option for the models is to update around every decade when major housing surveys are completed that directly measure electricity consumption. Individual and area consumption data will

reflect timepoints in society one decade apart. The process of releasing information is slow – the release of new data from both housing surveys and the UK census takes 3-5 years and the area classification system for the 2001 census was released for super output areas in 2007 after a full review of the methodology. The data for the year 2001 was represented by the fuel subsample of the 1996 English House Condition Survey, collected in 1997-99, and the 2001 UK Census was collected in 2001. In the future, the year 2011 will be represented by the energy follow-up survey to the 2009-10 English Housing Survey, collected in 2011-12, and the 2011 UK Census, collected in 2011.

5.4 Variable selection

This section will present a longlist of variables that are proposed to be applicable to the two main model options presented in the previous section. For each model option, a list will be compiled of the list of databases that are available for each identified purpose. All aggregate totals are for Lower Layer Super Output Areas, as they are the lowest level of aggregation by the Department for Energy and Climate Change's small area statistics on electricity consumption.

Table 5.2: Longlist of variables from the English House Condition Survey 1996

Short Code	Type	Name	Purpose
eannkwh	Continuous	Annual electricity usage (kilowatt hours)	To measure directly the non-heating end-use energy of a dwelling
ekwh1-9	Continuous	Electricity usage in each quarter (up to 9 quarters per case, over 2 years)	To measure the seasonal and annual variation of energy use
efirdate	Date	First date of measurement electricity use	As above
elstdate	Date	Last date of measurement electricity use	As above
chtyp96x	Categorical	Type of central heating	To determine if non-heating end-use energy equates to electricity use in the household
oheat96	Categorical	Other main heating provision	As above
floor96x	Continuous	Useful floor space (square metres)	To measure the physical size of the household
ten96x	Categorical	Tenure	To obtain a general classification of household socioeconomic status

Short Code	Type	Name	Purpose
goreg96x	Categorical	Government office region	To measure the regional dimension of energy usage
story96x	Interval	Number of levels above ground	To have a measure of intensity of the urban area
emphd96	Categorical	Employment status of household	To obtain a general classification of household socioeconomic status
type96x	Categorical	Type of dwelling	To measure the impact of housing typology
alinc96x	Continuous	Annual net income , all sources	To measure the impact of income on energy use
hhsize	Interval	Number in household	To measure the impact of the number of occupants
rooms96	Interval	Number of habitable rooms for exclusive use of households	To measure the physical size of the household

Table 5.3: Longlist of variables from the Living Costs and Fuel Survey 2008

Short Code	Type	Name	Purpose
gora	Categorical	Government Office Region	To measure the regional dimension of energy usage
accom	Categorical	Accommodation: Please code the household's accommodation	To measure the impact of housing typology
nrms, nrms2.. nrms6	Interval	Accommodation: How many of the following rooms do you have?	To measure the physical size of the household
numhhldr	Interval	Number of householders	To measure the impact of the number of occupants
occd	Categorical	Job: Occupation description	To obtain a general classification of household socioeconomic status
gwkinc	Continuous	Computed gross weekly income (dependent variable)	To measure the impact of income on energy use
centh	Binomial	Central Heating: Do you have central heating, including storage heaters, in this accommodation?	To determine if non-heating end-use energy equates to electricity use in the household

Short Code	Type	Name	Purpose
chfuel	Categorical	Central Heating: What fuel does it use?	As above
elecpay	Categorical	Electricity: By which of these methods do you pay for your electricity at this house or flat?	As above
eacamt	Continuous	Electricity: How much did you pay last time, excluding rental of appliances, hire purchase, loans or regular maintenance charges?	To determine the amount of the last bill for electricity
eacper	Interval	Electricity: What period did this cover?	To take account of seasonal and annual variations
ebbsamt	Continuous	Electricity: How much was your last budgeting scheme payment?	To determine the amount of the last bill for electricity
ebbsper	Interval	Electricity: What period did this cover?	To take account of seasonal and annual variations

Table 5.4: Longlist of census variables from the 2001 Census for Lower Layer Super Output Areas

Short Code	Type	Name	Purpose
UV60	Binomial	Amenities: with central heating	To determine if non-heating end-use energy equates to electricity use in the household
UV02	Continuous	Population Density	To have a measure of intensity of the urban area
UV61	Interval	Lowest floor level	To have a measure of intensity of the urban area
UV51	Interval	Number of People Living in Households	To measure the impact of the number of occupants
UV57	Interval	Number of Rooms	To measure the physical size of the household
KS16	Categorical	Accommodation Type	To measure the impact of housing typology
KS14	Categorical	Socioeconomic classification	To obtain a general classification of household socioeconomic status

Short Code	Type	Name	Purpose
UV63	Categorical	Tenure - Households	To obtain a general classification of household socioeconomic status

Table 5.5: Longlist of variables from the 2008 Department of Energy and Climate Change Small Area Statistics Database for Lower Layer Super Output Areas

Short Code	Type	Name	Purpose
	Interval	Number of domestic electricity meters	To measure directly the non-heating end-use energy of a dwelling
	Interval	Number of domestic economy7 electricity meters	As above
	Continuous	Consumption of domestic electricity meters	As above
	Continuous	Consumption of domestic economy7 electricity meters	As above
	Continuous	Consumption of domestic gas meters	As above

Table 5.6: Longlist of variables selected for the 2001 Area Classification for Super Output Areas

Short Code		Name	Purpose
UV02	Continuous	Population Density	To have a measure of intensity of the urban area
UV60	Binomial	Amenities: with central heating	To determine if non-heating end-use energy equates to electricity use in the household
UV61	Interval	Lowest floor level	To have a measure of intensity of the urban area
KS19	Continuous	Average household size	To measure the physical size of the household
KS15	Categorical	Travel to Work	To have a measure of intensity of the urban area
KS16	Categorical	Accommodation Type	To measure the impact of housing typology
UV58	Continuous	Persons per Room - Households	To measure the impact of the number of occupants

Short Code		Name	Purpose
KS12	Binomial	Unemployed	To obtain a general classification of household socioeconomic status
KS14	Categorical	Socioeconomic classification	To obtain a general classification of household socioeconomic status
UV63	Categorical	Tenure - Households	To obtain a general classification of household socioeconomic status

5.4.1 Annual option variables available

Table 5.7: Summary table of annual option variables that fulfil the recommendations for the inclusion of data

	Individual		Area			
Purpose	**EHCS**	**LCFS**	**CEN**	**SAS**	**CLASS**	**Notes**
To measure directly the non-heating end-use energy of a dwelling	■	■		■		1
To measure the seasonal and annual variation of energy use	■	■				2
To determine if non-heating end-use energy equates to electricity use in the household	■	■	■		■	1
To measure the physical size of the household	■	■	■		■	1
To obtain a general classification of household socioeconomic status	■	■	■		■	1
To measure the regional dimension of energy usage	■	■	■	■	■	1
To have a measure of intensity of the urban area	■		■		■	3
To obtain a general classification of household socioeconomic status	■	■	■		■	1
To measure the impact of housing typology	■		■		■	3
To measure the impact of income on energy use	■	■				2
To measure the impact of the number of occupants	■	■	■		■	1

Abbreviations used in 5.4.1 and 5.4.2

EHCS	1996 English House Condition Survey
LCFS	2008 Living Costs and Food Survey
CEN	2001 Census
SAS	2008 Department for Energy and Climate Change Small Area Statistics
CLASS	2001 Office for National Statistics Area Classification for Super Output Areas and Data Zones

5.4.2 Decennial option variables available

Table 5.8: Summary table of decennial option variables that fulfil the recommendations for the inclusion of data

	Individual	Area			Note
Purpose	**EHCS**	**CEN**	**SAS**	**CLASS**	
To measure directly the non-heating end-use energy of a dwelling	■		■		1
To measure the seasonal and annual variation of energy use	■				2
To determine if non-heating end-use energy equates to electricity use in the household	■	■	■	■	1
To measure the physical size of the household	■	■		■	1
To obtain a general classification of household socioeconomic status	■	■		■	1
To measure the regional dimension of energy usage	■	■	■	■	1
To have a measure of intensity of the urban area	■	■		■	1
To obtain a general classification of household socioeconomic status	■	■		■	1
To measure the impact of housing typology	■	■		■	1
To measure the impact of income on energy use	■				2
To measure the impact of the number of occupants	■	■		■	1

Notes

1 – Both individual and area level data available. Model can use either outcome or predictor variables; aggregate data can be used in verification and future refinements.

2 – No area-level data available. The individual-level data is useful to explore correlations and to estimate consumption levels in 2008 for homes covered in the fuel sub-sample conducted between 1996 and 1998 in the annually updated model, but these variables are unsound for use in the models that include area-level variables.

3 – Information available only in annually updated datasets. This is not necessarily advantageous as these factors typically change very slowly.

The most notable omission in this process of variable selection is the removal of income and seasonal and annual variations as predictor variables in the final model in both options. The correlations between growth in income in real terms and non-heating end-use energy consumption at the macro-scale has been documented in previous social research, both in long-term trends and short-term price sensitivity (Office for National Statistics, 2010, Lenzen et al., 2004, Nesbakken,

1999, Summerfield et al., 2007). The role of seasonal and annual variations in non-heating end-use energy has less of an evidence base. Variable seasonal pricing is used for heating fuels such as natural gas by the British power industry, but not for electricity, which is less likely to be used as a heating fuel (Npower, 2011, British Gas, 2011, Department of Energy and Climate Change, 2009a). The current functions estimating the fluctuation of non-heating energy usage are based on the results of the 1996 English House Condition Survey, but are currently deemed experimental (Smith, 2011, Sefton and Chesshire, 2005).

The result is a range of variables in large, designed surveys and aggregate statistics that contain information about non-heating end-use energy consumption that are open and available to the researcher for investigation. The following section will detail the methodological options considered, from methods that only use aggregate data, to those that only use individual entries in a horizontal structure, and those that use individual entries in a hierarchical structure.

5.5 Range of applicable quantitative methodologies

This section will outline the range of quantitative social science methodologies available for investigating non-heating end-use energy. In this research, domestic energy use modelling is taken to be the outcome, or dependent, variable with all other variables as predictor variables. Domestic energy use modelling currently estimates the energy use of individual households using data collected from the entire residential sector. Generally, the primary purpose of domestic energy modelling is driven by the heat balance, or the difference of the external and internal temperature, of a dwelling and not for the direct estimation of non-heating end-use energy. The new availability of area classification data for housing surveys presents new methodological options that can be considered for estimating non-heating end-use energy. This section will present these methods and evaluate their suitability for the direct estimation of non-heating end-use energy of households that occupy both new buildings and existing buildings.

There were seven methodologies that were considered for use as either the annually updated or the decennially updated model options. These were econometric, technological, ecological, multiple regression, archetypal, growth, and multilevel models. Each of them have different strengths and weaknesses when satisfying the criteria set out for the two model options because the frequency of data collection for energy consumption as part of a national housing survey is sporadic, the classification of areas in England is made every ten years, but collection of expenditure surveys and aggregate totals of energy use and the number of meters used in the domestic sector are made annually.

Conditional demand analysis was explored in detail in Chapter 3 as the framework for previous estimates of non-heating end-use energy in the heat balance equations used as part of the building regulations in England using technological, multiple regression, and growth models. To recap:

$$HEC_{it} = \sum_{j=1}^{j} UEC_{ijt} \times S_{ij} \quad (28)$$

where HEC_{it} was presented as the total non-heating end-uses energy consumption by household *i* in period *t*, UEC_{ijt} is the *j*-unit end-use energy consumption of household *i* in period *t*, and S_{ij} is a binary predictor of household *i*'s ownership of device *j*.

Unit Energy Consumption UEC_{ijt} was presented as a function of the features of every household *i*'s energy consuming unit *j* (AF_j) and the utilisation pattern UP_{ijt} that relates to energy-using device *j*. The utilisation pattern of the device itself was presented as a function of the structural aspects of the dwelling ST_i, weather conditions WC_{it}, market conditions MC_{it}, and the socioeconomic situation of the household SEC_i. This results in a function of UEC_{ijt} as

$$UEC_{ijt} = F_j(ST_i,\ AF_j,\ WC_{it},\ MC_{it},SEC_i) \quad (29)$$

The work in Chapter 3 showed how this perspective can be changed to model households as collections of different types of end-uses *j*, such as appliances and electronics, lighting, and cooking, that all use electricity instead of attempting to obtain appliance and electronics ownership patterns and modelling the electricity use of every non-heating energy-consuming device. In this case, the "ownership" variable S_{ij} is stable for the number of different kinds of end-uses that are measured in the model. For non-heating end-uses, this number is three in the current SAP2009 model (appliances, lighting, and cooking). For the purposes of creating a model out of survey data from houses that only have electricity consumption data that equates to all non-heating end-uses, the number is only one.

As discussed in Chapter 3, housing survey data used to calculate the baseline consumption of energy use does not contain information on the ownership of low energy lights or cooking appliances. Correction factors around low-energy lighting are derived from tests outside of the evidence base of housing conditions surveys, and cooking appliance assumptions are derived from statistics from the Market Transformation Programme (Market Transformation Programme, 2008, Henderson, 2009). Therefore, unless stated, most of these models will assume that $HEC_{it} = UEC_{ijt}$ with all homes modelled on the basis of consuming natural gas for space and water heating and electricity for non-heating.

5.5.1 Top-down statistical options: Econometric and technological models

Domestic energy modelling can be conducted to estimate the overall energy use of the residential sector. A range of methods are available and have been used from econometric-style "top-down" methods to archetypal "bottom-up" methods in what is commonly named housing stock modelling. Top-down methodologies are appropriate when the data is only available in aggregate form and the data has been available for an extended period of time with no periods of discontinuity. Two main branches have been described by previous authors as *econometric* and *technological* modelling (Swan and Ugursal, 2009).

Econometric modelling involves the use of a "theory-based forecasting model" that operates at the macro-level and takes in cross-sectional or time-series data (Greene, 2008, Angrist and Pischke, 2009). In the language of the conditional demand framework,

$$UEC_{ijt} = F_j(ST_i,\ MC_{it}\ , SEC_i) \qquad (30)$$

where HEC_{it} is a measure of the central tendancy of household non-heating energy consumption for all households *i* in a defined geographic area (e.g. a nation, region, or area) during period t, ST_i is the central tendency of household size in numbers of people, MC_{it} is a measure of price elasticity (the change in demand per change in price), and SEC_i is a measure of the socioeconomic status of a household.

This type of model is extremely easy to update using yearly data if available, as it does not rely on occasionally-gathered housing survey data that includes energy consumption. However, it is not a reliable model for predicting the energy consumption of a single household because of the effect of aggregation of all households, resulting in a narrower distribution of energy use than in reality. This type of model would not be viable using area-based variables because of its reliance on economic, rather than physical, data about households. Area-level data about income is not collected from the entire population in the census, therefore the aggregate statistics are created from calculations from other models, such as the Experian Mosaic model at Lower Layer or the Office of National Statistics at Ward Level. This methodology is useful for forming hypotheses about patterns of overall behaviour relating to non-heating end-use energy, but not for individual households.

Technological modelling involves constructing a conditional demand analysis model as the national scale. A classic example was discussed earlier, DECADE, that was implemented in the UK and adopted by the researchers formulating the BREHOMES housing stock model for non-heating energy in households from 1996 to 2010 (Environmental Change Institute, 1995, Shorrock and Dunster,

1997). This type of model does depend on the patterns of ownership of appliances which means $HEC_{it} \neq UEC_{ijt}$. Therefore, technological modelling is defined by:

$$HEC_{it} = \sum_{j=1}^{j} F_j(AF_j) \times S_{ij} \quad (31)$$

where AF_j is the appliance features of each energy-consuming device *j* where the rate of consumption and assumed number of hours of use depend on the vintage, or year made, of the appliance. This model assumes that energy use in dwelling *i* is determined by the stock of appliances in that dwelling.

Again, this model is intended for updating on a yearly basis by estimating the number of energy-consuming devices installed in households for non-heating purposes. Aggregations of the number of appliances are estimated using proprietary data and data collection is unharmonised and highly fragmented amongst the professional and trade bodies covering different non-heating end-uses such as wet appliances, electronics, telecommunications equipment, and cooking. For lighting, an area that proposes significant correction factors for low energy light bulbs, no overall data existed on sales or ownership of these or any type of light bulb according to the researchers involved with DECADE (Environmental Change Institute, 1995). Therefore, a technological model is useful, and has been used in BREHOMES, as a model that estimates the future use of the housing stock.

This thesis does not propose to bring forward econometric or technological methods for use in this research. This section summed up the benefits and drawbacks of these methods. Using these models in the annual modelling option is beneficial for ease of use and availability of information. However, the information available in aggregate does not cover the entire population, but instead depends on a second model to supply this information in the background. Finally, such models are useful for alluding to and describing general trends in society that relate to non-heating end-use energy, but the use of these methods to model single households would potentially create additional errors. These errors would arise for an incorrect distribution of "central tendencies" derived from aggregate data that would be narrower than for the true population.

5.5.2 Bottom-up statistical methods: Archetypal method

Archetypal methods have been in use in housing stock modelling as a way to model the residential sector's energy use, but have not been explored as models of individual dwellings. These methods develop a collection of "typical" households that are modelled as individuals and then multiplied by a grossing factor to estimate the energy use of the national population of households. These typical households can either be a dwelling category, defined by broad characteristics of household type, size, heating method, and other data collected in housing surveys, or they can be an archetype, a

carefully constructed physical model of a building that includes all the architectural and engineering details of the dwelling. These details can include building materials, floorspace, occupants, orientation, and room configuration. Researchers have created typical households of around 50 archetypes to numbers of dwelling categories that have ranged from 1,000 to 20,000 (Kavgic et al., 2010, Parekh, 2005). For heating end-uses, the loss of detail is a drawback for collections of dwelling categories, but this is not the case for non-heating end-uses as architectural and engineering details are not necessary for estimating non-heating end-use consumption figures. The energy use of a household is then entirely defined by its archetype defined by a series of categorical variables instead of a regression equation with predictor interval or continuous variables.

For non-heating end-uses, the archetypal method only needs details about the type and size of the building, providing that the method only considers the super-majority of homes in England that use different fuels for heating and non-heating. However, as detailed in Chapter 3, the housing stock model assigns higher energy usage for non-heating in different dwelling categories. These categories are assumed to belong to different socioeconomic classes. Therefore, the archetypal model that has been in use England in the recent past, specifically in BREHOMES from 1996 to 2011 (Palmer, 2010) is:

$$UEC_{ijt} = F_j(ST_i, SEC_i) \quad (32)$$

Where ST_i represented the size of the dwelling measured in usable floor space, and SEC_i is a scaling factor assigned to the "top" and "bottom" slices of the housing stock, as explored fully in Chapter 3.

Archetypal models are potentially valuable in conjunction with the decennial and annual options. Housing surveys are collected annually at the national level, but energy consumption from meters is only collected occasionally. Census data contains the housing categories for the entire population, and area classification data relies on data on housing categories similar to those used in bottom-up housing stock models built from single dwelling models of archetypes. Housing surveys and census data include variables for the number of rooms, housing type, and number of occupants. Therefore, in the decennial option, the measure of household size should be the number of rooms instead of the amount of usable floorspace to maintain data harmonisation. The dwelling categories are going to be more harmonised in the future, as the 2011 census has recently included heating fuel as well as the availability of central heating in the questionnaire.

There are three main weaknesses of the archetypal method. The first is that the main purpose of the design of the method in the past has not been to estimate the energy consumption of a single dwelling, but the entire energy consumption of the dwelling stock. The second weakness is that in

order to apply housing survey data designed for the national scale to an individual scale, the constancy assumption needs to be invoked. A constancy assumption in this case would assume that housing types would have the same proportion of small, medium, and large numbers of occupants inside them and vice versa. This is because the exact details of occupancy of individual homes are not available for the entire population in small census areas (e.g. counts are available for one-person households and two-room households, but not for one-person/two-room households). A third weakness is that in the future, the refining of the archetypal model with actual consumption data from small area statistics will reintroduce scaling factors to reduce the variance between predicted and actual non-heating end-use energy consumption.

The archetypal method will be brought forward for more detailed examination in this thesis. The method has usefulness in the wide availability of data on housing types and sizes from both housing surveys and census data. However, there are flaws in the archetypal method that are likely to prevent its use as a model of single household non-heating end-use energy involving constancy assumptions and scaling factors as more advanced methods dealing with both individual level and area level predictors are developed in the future.

5.5.3 Using only aggregated data – ecological method

An alternative approach to the estimation of household energy use is to ignore housing surveys that contain energy consumption data, and instead rely solely on aggregate statistics that cover the entire population. In this method, the researcher obtains an average energy consumption figure per household in every neighbourhood with corresponding predictor variables of, for example, the average size of households in that area. The use of such an approach brings the risk of committing the ecological fallacy (Robinson, 1950, Robinson, 2009), or the interpretation of group results as if they apply to an individual dwelling. Finding relationships between the averages of variables instead of using individual-level data and samples therefore is risky. The strength of the method is that if the right data is available, the entire population can be covered with much less non-response and self-selection biases present in housing surveys, especially when derived from voluntary programmes designed to save energy or for regeneration and housing renewal (Hartman, 1988). In contrast, participants in the census and in the collection of aggregate energy data cannot refuse to take part, as the census questionnaire in the United Kingdom is mandatory and the energy balancing and settlement regulations collect energy data automatically (Statistics Commission, 2007, Elexon, 2010).

Like the archetypal method above, the ecological method could encompass the household size and socioeconomic variables of households for non-heating end-use energy as a single end-use *j*:

$$UEC_{ijt} = F_j(ST_i, SEC_i) \quad (33)$$

The ecological method can apply to the decennial option for estimating non-heating end-use energy. An annual approach will require assumptions of little change over the course of a decade in the population, as data that covers the entire population from the census is only available every ten years in the United Kingdom, but data regarding overall energy consumption is available annually for each super output area in England. The main strengths of this method are that the entire population is covered, but there are significant weaknesses. The first weakness is that averaging tends to dampen extreme values, and the predictive power of small or large homes, for example, would be likely to be lost when creating an ecological model with continuous dependent variables. The second weakness is that the model would have to rely on the "constancy assumption", or the assumption that an individual household's behaviour in the use of non-heating end-use energy does not depend on geographic location due to the combining of the average values of a number of different areas (Freedman, 1999).Therefore, difference between neighbourhoods and area classifications would be lost using the ecological method.

5.5.4 Using only individual level data – linear regression and growth model methods

However, if the purpose of the outcome is to estimate the energy use of an individual household, the methodological options are more limited to the area of regression analysis. If no individual household level data is available, ecological regression techniques can be appropriate, but the likelihood of making a Type I or Type II error is extremely high. Instead, the approach taken in current modelling of non-heating energy end-uses in England has used linear regression modelling as the predominant method. There have been two major iterations of the linear regression model since algorithms were introduced to replace lookup tables in the mid-1980s. The first of these was the proposal for a combination of quadratic and linear regression equations to represent non-heating energy consumption, and the second was a growth model equation. These are both based upon classic statistical approaches to linear modelling involving data that is assumed to be approximately parametric.

The general function of these approaches only considers the size of the household, defined by usable floor space for non-heating end-uses *j* (lighting, appliances and electronics, and cooking):

$$UEC_{ijt} = F_j(AF_j,\ ST_i) \quad (34)$$

The polynomial and linear regression models were first created by Henderson, Shorrock, and Chapman as part of the early development of BREDEM in the 1980s. The lastest version, using from the publication of BREDEM-8 and BREDEM-12 in 1996, was

$$E_{LA} = 619 + 6.44\ (TFA \times N) \text{ if } TFA\ \times N < 710 \quad (35a)$$

$$E_{LA} = 2700 + 4.05\ (TFA \times N) - 0.214\ (TFA \times N)^2 \text{ if } 710 \leq TFA\ \times N < 2400 \quad (35b)$$

$$E_{LA} = 7990 \text{ if } TFA\ \times N \geq 2400 \quad (35c)$$

where E_{LA} is the electricity consumption for appliances and lighting in kilowatt-hours per annum, *TFA* is the total floor area in square metres, and *N* is the calculated number of occupants dependent on floor area (Anderson et al., 1996). These algorithms are baseline figures with correction factors from the physical attributes of low-energy lightbulbs and the daylighting of the building represented by lighting end-uses AF_j and structural attributes of the building ST_i.

The growth model for non-heating end-use energy was introduced by the latest iteration of BREDEM in the 2009 revision of the Standard Assessment Procedure. The model for lights and appliances (BRE, 2010),

$$E_i = 267.53 \times (TFA \times N)^{0.4714} \quad (36)$$

was an initial baseline figure for energy use due to lights and appliances in kilowatt-hours per annum with correction factors based on seasonality and on the physical attributes of low-energy lightbulbs and the daylighting of the building.

5.5.4.1 Multiple independent variables in linear regression

All of these algorithms used data from housing surveys collected as a representative, often weighted, sample of the entire population, and use a linear regression equation to model the amount of non-heating end-use energy in the building as represented by electricity use when heating used a different fuel which almost always was natural gas. Regression is the main statistical technique used to explain the relationship between a dependent variable, in the case of the thesis, non-heating end-use energy use in residential buildings, and any number of independent or explanatory variables.

A linear regression analysis is the fitting of a straight line to the scatterplot of Y, the dependent variable, against X, the independent variable(s). This equation is

$$y = mx + c, \text{ or alternatively } y = \beta_0 + \beta_1 x \quad (37, 38)$$

The slope *m* or β_1 is the slope coefficient of the independent variable *x* and *c* or β_0 is the intercept or the point at which the line cuts the Y-axis. For any given individual household *i* from 1 to *n* the linear relationship between Y and X is the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (39)$$

Where e_i is the residual, or the difference between the actual y-value and the one that is predicted by the x-value. All the residuals represent the "scatter" of points about the regression line. There are special assumptions that must be made about the residuals, known as the parametric tests. First, the residuals are assumed to have a normal distribution with a mean of zero with a variance noted as σ^2. Second, the residuals are homoskedastic, or the variance of the residuals remains constant for any range of *x* chosen. Third, the residuals are not correlated with each other. If these assumptions are not met, the estimates of β_0 and β_1 might be proven to be biased.

In linear regression, the estimates of β_0 and β_1 are estimated from the observations using the least squares method that minimises the sum of the squared residuals. These are denoted by the symbols $\hat{\beta}_0$ and $\hat{\beta}_1$ representing the slope and intercept coefficients. Using this method, the researcher can obtain the fitted regression line that predicts a value for *y* for an individual *i*:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x \quad (40)$$

Even in the situation where the researcher is interested in only one dependent variable, he needs to take account of other variables as they may compromise the results. The situations include:

- Inflation of the relationship, for example between electricity use and floorspace as the apparent strong relationship could diminish
- Suppression of the relationship. The apparent relationship that is weak could increase
- Appearance of no confounding in the relationship before more variables are introduced

There are some steps to the teasing out of complex relationships between a dependent variable and multiple independent variables, often called multiple regression:

- The first step is the make a quantitative assessment of the size of the effect of a variable – for example, the effect of floorspace on electricity use
- The second step is the make this same quantitative assessment after taking account of other variables – for example, taking account of the connection between household income on floorspace
- The third step is the state a measure of the size of the uncertainty of the original effect (floorspace)

The traditional method of dealing with more than one independent variable is to give each of them its own term *x*. The regression equation using individual household data with two independent variables would then be

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad (41)$$

where y_i is the outcome variable of non-heating end-use energy in a household, β_0 is the intercept, β_1 and $\beta_2$are two slope coefficients relating to independent variables x_{1i} and x_{2i}.

However, current and past modelling of non-heating end-use energy does not use more than one term (x-variable) to represent the independent variables of total floor area and numbers of occupants. Instead, the interaction effect between the two variables is used $\{TFA \times N\}$. An interaction effect is "the differing effect of one independent variable on the dependent variable, depending on the particular level of another independent variable" (Cozby, 1997). This is because it was claimed by modellers of domestic energy use in the 1980s that both indicate increased consumption, but an increase in occupants multiplies the effect of increased floor space in residential buildings.

In multiple regression, the effect of one independent variable may be more pronounced when taking account of another independent variable. This is known as an interaction effect on the dependent (Y) variable when the effect of one of the independent variables depends on the value of another independent variable.

The data input to the multiple regression can be from different types of independent variables. The optimal situation is when there are all continuous variables involved, but in research is it often necessary to use interval or categorical variables. It is not possible to perform a multiple regression where the dependent variable is not continuous. The two independent variables chosen for analysis out of the datasets were total floor area, which was a continuous variable measured in square metres, and the numbers of occupants, which in this context is an interval variable. An interval variable represents data measured on a scale where all the intervals are equal along its entire length. In other words, an increase from two to four people always will have double the effect.

For example, in a two-variable multiple regression where x_1 is continuous and x_2 is an internal variable $\{0,1,2,\cdots n\}$, the resulting lines for each case of x_2 are the same slope with intercepts $\{\hat{\beta}_0, \hat{\beta}_0 + 1, \hat{\beta}_0 + 2, \cdots \hat{\beta}_0 + n\}$. Is it reasonable to conclude that the difference in *y* between the intervals is the same for all instances of x_1?

There are some methods available to provide an alternative conclusion. The most obvious is to split the sample into each category and fit a regression for each category of household size. However, there are several fallacies that can occur. The sample size can become unacceptably small. There can be more than one categorical predictor, and the effects of other independent variables can vary for each possible group. There can be several independent variables predicting the dependent variable, but not all of these variables will vary across the grouping, leading to redundant and separate regressions. Finally, and crucially, hypothesis tests cannot be carried out to compare regression coefficients across the different regressions for each grouping.

One solution, and this has been implemented in the BREDEM / SAP model currently in use in the United Kingdom, is the fitting of a pooled sample as the product of the two independent variables. The pooled variable $x_1\_x_2 = x_1 \times x_2$ in this case replaces the previous two-variable multiple regression

$$y_i = \beta_0 + \beta_1 x_1\_x_{2i} + e_i \quad (42)$$

$x_1\_x_2$ is known as the interaction effect between x_1 and x_2 and allows the effect of x_1 to differ for each interval of x_2. An interaction effect is found if the effect of x_2 multiplies the effect of x_1.

This interaction effect therefore declares that if there are additional occupants inside a dwelling, it multiplies the effect of additional floorspace. In the context of non-heating end-use energy, this is a plausible hypothesis: buildings do not use electricity without demand from people, and people need devices that "inhabit" buildings with them to operate. The causes of non-heating energy use involves human activity in an environment that offers opportunities for energy use, together with energy use that is a function of the ownership of appliances themselves. It would be reasonable to expect a causal model to include measures both of the number of occupants in a dwelling and some approximation of the number of appliances expressed as physical household size, probably as a product. This is a tentative qualitative explanation for the sturdiness of the interaction term in algorithms since the 80s.

5.5.4.2 The benefits of the interaction term TFAxN

The evidence bases that gave rise to the BREDEM-8/12 and the SAP2009 algorithms for estimating the annual household consumption of baseline non-heating end-use energy had a feature that was proving troublesome for linear regression modelling. This feature was a dampening of the effect of increasing household size on these end-uses. As household sizes got bigger, the amount of energy they use tended to flatten out as if it belonged to a curve with a horizontal asymptote.

With the use of a single independent term instead of two separate terms, there are options to use linear regression techniques to do non-linear modelling. Polynomial regression, used in BREDEM-8/12 and growth models, used currently in SAP2009, are all linear from the point of view of estimation. There are other options available for non-linear regression with traditional non-parametric methods and bootstrapping. However, these methods are weaker and more labour-intensive than parametric statistical methods.

Non-parametric methods are able to overcome the situation where the data is not approximately parametric. Two examples of these methods, the Mann-Whitney U test and the Wilcoxon Signed-Rank test, convert a dataset with an interval or continuous dependent variable to a ranked list of cases to assess. The rationale is that the differences between the raw scores are alternatively small and large, and these differences even out over the course of all scores. However, converting data with an interval or continuous dependent variable to ordinal ranked data results in a loss of sensitivity. Sensitivity is the power to reject the null hypothesis, or the starting point that differences do not occur in the population, and therefore lower sensitivity gives a higher type II error rate. As this thesis follows the prevailing theory that non-heating end-use energy is dependent on household size, the risks of type II errors are inflated (Uglow, 1982, Henderson and Shorrock, 1986b, Environmental Change Institute, 1995, Henderson, 2009, Energy Advisory Services, 1996).

Bootstrapping is defined by repeated randomised sampling of cases from a non-parametric dataset. The number of cases in a subsample will be equal to the number of cases in the original dataset, meaning that some cases can be randomly selected more than once. This process is repeated a number of times with a correlation coefficient calculated for each subset against the dependent variable. The original correlation between the independent and dependent variables obtained using linear regression methods could be trustworthy if there is a normal distribution around an average close to the original correlation coefficients of all the subsamples. This method is still in its infancy and how far the method can be applied is still indeterminate (Wilcox, 2005). This method is noted to be weak in its application to skewed data that is commonly found in energy data (Department of the Environment Transport and the Regions, 2000b, Department for Communities and Local Government, 2006b, Kaza, 2010). Domestic energy data that depends on household size will always have a “long tail” of larger dwellings. Dwellings with eight or more rooms took up 11% of the total housing stock in England according to the UK census compared with 4% of homes with less than three rooms (Office of National Statistics, 2005).

The creators of the non-heating end-use prediction algorithm in BREDEM-8/12 used a combination of straightforward linear regression and polynomial regression. For homes where the interaction of

total floor area and numbers of occupants is less than 710, the model only required simple linear regression. Using homes with valid energy consumption data from the fuel subsample of the 1996 English House Condition Survey, 97.5% of households in England were in this category. A two-term polynomial regression model was fitted to most of the remaining 2.5% of English households. The polynomial simply adds the interaction term, squared, as a second independent variable. This creates a multiple regression equation

$$y_i = \beta_0 + \beta_1 x_{1\_} x_{2i} + \beta_2 (x_{1\_} x_{2i})^2 + e_i \quad (43)$$

Similar assumptions are made in multiple regression as is made in single regression – first, that the residuals are normally distributed, and second, the variance of the residuals are homoskedastic, meaning residuals are not correlated with each other. In multiple regression, residuals are standardised, or having a mean of zero and a standard deviation of 1 to place the residuals of all independent variables on the same scale.

However, the use of linear regression techniques for the modelling of non-heating end-use energy is curious considering that the research team involved with the validation of BREDEM stated that the data they had did not meet the parametric tests (Shorrock et al., 1991, Shorrock et al., 1994). These papers given to the 1991 and 1994 conferences of the Building Environmental Performance Club state that although the data available for appliances, lighting, and cooking do not meet these statistical tests, their recommendation was to proceed with the linear regression model because its presence is better than a situation where no model is available to estimate the heat generated (Energy Advisory Services, 1996, Shorrock and Dunster, 1997).

In the opinion of this researcher, this lack of concern was because BREDEM and SAP were designed and functioned as a heat balance equation, not an energy consumption model. The heat levels generated are very low compared with heating end-uses, and consequently the risk posed by errors, in what was considered to be a second-order correction, was deemed to be low. On the other hand, it would have been damaging for the credibility of the project team to government funders to have missing the types of consumption that people interact with more regularly than heating controls, such as cooking or turning on lights and the television. When SAP was released to be used in the building regulations in the 1990s, the rating was based entirely on the running costs of heating end-uses as required by Part L of the building regulations for the conservation of fuel and power (Department of the Environment, 1995). As a consequence, this admittedly unreliable equation was kept as part of the heat balance equation.

After criticism of the treatment and reliability of non-heating end-uses in this model in internal and external review (Sefton and Chesshire, 2005, Henderson, 2009), a revision to the baseline algorithm was made based on housing survey data in the fuel sub-set of the 1996 English House Condition Survey instead of survey data generated within the buildings research community. The new algorithm still used the interaction of two independent variables, total floor area and the number of occupants, as a single term without any review of this practice. The combination of linear regression and polynomial regression was replaced with a growth model for E_L electricity consumption due to lights and E_A electricity consumption due to appliances, cooking, and electronics

$$E_L = 59.73 \times (TFA \times N)^{0.4714} \quad (44)$$

$$E_A = 207.8 \times (TFA \times N)^{0.4714} \quad (45)$$

where the equation takes the form

$$\hat{y} = a\hat{x}^b. \quad (46)$$

A growth model is based on a linear equation of the transformation using a natural logarithm of electricity use ($E_L + E_A$) and of the interaction term. The data in this model was, as will be shown in detail in Chapter 6, positively skewed, meaning that there was a "long tail" of larger households and larger energy-consumption households in the dataset that resulted in the data failing the parametric test. One method of using linear regression on the data is to take the natural logarithm of both the dependent variable and the interaction term encompassing both independent variables and subject them again to the parametric tests. Because the growth model is in SAP2009, one can presume that these tests were passed, meaning that the logarithmic transformation of both terms are approximately normally distributed, or "log-normal." This linear regression equation using transformed variables was solved as

$$\ln(E_L + E_A) = \ln(59.73 + 207.8) + 0.4714 \ln[TFA\ x\ N] \quad (47)$$

Because of the mathematical properties of natural logarithms, it is possible to back-transform these transformed variables to the form $\hat{y} = a\hat{x}^b$ to arrive at the two equations for lights and appliances above. The split between lighting and the other non-heating end-use was necessary to allow correction factors for daylighting and low-energy light bulbs. The evidence used for the split of lights and appliances comes from end-use monitoring in a more limited number of dwellings (Henderson, 2009).

The growth model in SAP2009 solved two problems by using a linear regression equation in BREDEM-8/12. The first of these problems was an unwieldy combination of three equations depending on the size of the household (one of these equations is a constant). The second was that underestimation of non-heating energy use was occurring because the levelling-off of energy use in relation to household size in the housing survey data was occurring well before the 97th percentile of homes, making the linear equation ill-equipped to accurately estimate consumption.

5.5.4.3 Applying linear regression techniques to the annual and decennial options

There are still good opportunities for the application of linear regression techniques to both the annual and decennial options presented earlier in this chapter. Because housing survey data that includes energy consumption data is collected rarely, this type of approach is most valid for the decennial option. However, use of related annual surveys that contain some data on energy expenditure per household and indications of household size may allow use of these techniques in the annual option. One current survey, the Living Costs and Food Survey, surveys the number of rooms instead of floorspace and the last quarterly or monthly bill instead of requiring more than 9 months' data as in the 1996 EHCS.

There are benefits and drawbacks to using linear regression techniques for estimating non-heating end-use energy of single households. If the data can be modelled using linear regression techniques, either in raw form or transformed in the creation of a growth model, it would be easily understood by most of the building research community. It would also be possible to use the longstanding interaction terms of floorspace and numbers of occupants, but that is also a drawback of previous applications of linear regression methods because of the exclusive use of these two terms to create a physically-derived model. This model uses the interaction term to the exclusion of other variables that may add to its accuracy. One variable that has been proposed is the type of area in which the household resides; as a categorical variable, it would be problematic to include in a dataset where the dependent variable and the interaction term representing two independent variables are both log-normal, resulting in a growth model to the exclusion of other independent variables. The following section, covering multilevel methods, puts forward a proposal to classify the cases by area instead of assigning a new independent variable to each case.

5.5.5 Using classed individual level data and aggregate data – multilevel methods

5.5.5.1 Introduction

The multilevel method is a way to explore whether non-heating end-use energy outcomes vary across areas and to explore the effect of living in a certain area on an individual household. It does not assume that all households are making decisions about non-heating end-use energy

independently. Instead, the technique delves into the reasons that households agglomerate in neighbourhoods with varying levels of homogeneity. The thesis will examine a multilevel method using groups, multilevel linear modelling, assuming the data passes the parametric tests.

The general function of these approaches again considers the characteristics of the household, but also the same characteristics of its area classification *j* by not considering non-heating end-uses separately. The unit electrical consumption for these households and areas that do not use electricity for heating end uses, is:

$$UEC_{ijt} = F_i(CL_j, ST_i) \quad (48)$$

where ST_i represents any characteristic of the household, and CL_j is the mean of that characteristic for its area classification.

If one only uses individual level data to estimate non-heating end-use energy, this assumes that all households are independent of one another when it comes to the use of energy. There are several reasons why this assumption might not hold: groups of people are exposed to different energy-saving messages, the housing and urban structure of certain types of areas encourage different living patterns and styles of non-heating energy use, or households connected through social networks displaying different patterns of appliance and electronic device ownership. This thesis proposes that general-purpose area classifications are a useful way to approximate the interaction of these types of "unknowables" and "unquantifiables" for a national population.

Multilevel models enable the researcher to investigate the nature of the variance between groups, and the effect of group-level characteristics on outcomes at the individual level. When individual households form groups, one can expect that two randomly chosen individuals from the same group will tend to be more alike than two individuals chosen from different groups. These groups form over a long period of time, but there is evidence that households are motivated by the quality of the neighbourhood when locating, when there is no evidence that the new dwelling is better. These findings were most prevalent in the types of households that are expected to grow fastest in the future, including elderly one-person households and privately rented properties (Clark et al., 2006, van Ham and Clark, 2009, Feijten and van Ham, 2009).

Clustering of a particular energy-related issue can lead to the clustering of information campaigns and the efforts to "nudge" targeted populations into taking action. One example to illustrate clustering related to domestic energy use is the campaign to reduce fuel poverty. Information on the clustering of fuel poverty in small areas is used as part of the formulation of policy responses and

supplying fuel-saving and increased information on energy efficiency in the home (Department of Energy and Climate Change, 2010). For example, clustering of fuel poverty in rural areas was reported in Warwickshire (Warwickshire Observatory, 2011) and urban, multicultural areas in London (Association for the Conservation of Energy et al., 2008). The clustering of fuel poverty in certain types of areas in England, as defined by the ONS 2001 Area Classification for super output areas in England is outlined in Figure 5.1 below.

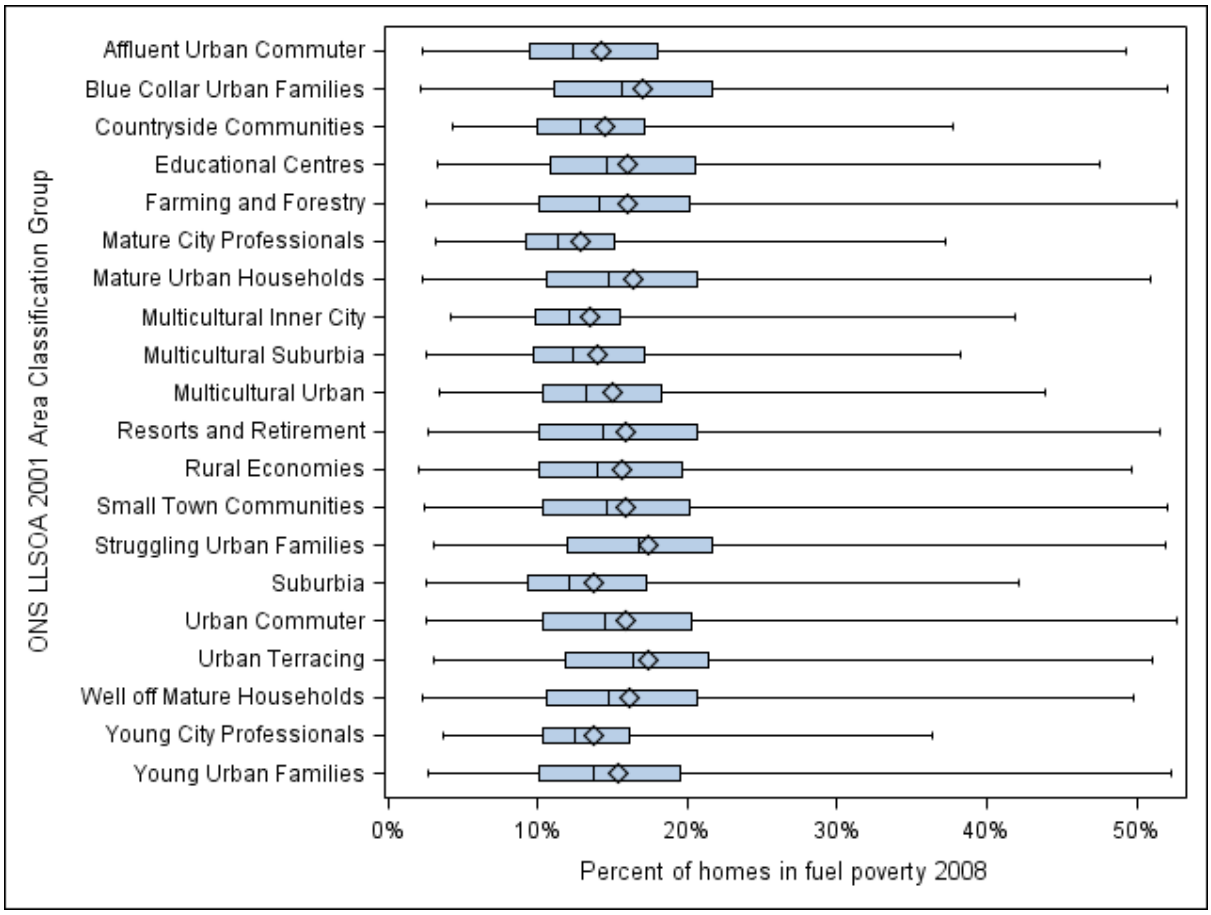


Figure 5.1: Fuel Poverty in England by area typology

The implications of ignoring clustering can led to conclusions that a group predictor of energy use is real when the effect is actually due to chance. This kind of error, where one predicts an effect where one does not exist in reality, is a Type I error. In a multiple regression equation, the standard errors associated with regression coefficients of group-level independent variables are generally underestimated. This underestimation can be corrected if variation between groups, in this case, between area classifications, is allowed instead of just assigning a group-level characteristic - for example, the population density of the super output area - to each individual household.

5.5.5.2 Multilevel models and variance within and between groups

To understand the structure of a multilevel model, one can begin by applying the principles to a model of non-heating end-use energy without any explanatory variables:

$$y_i = \beta_0 + e_i \quad (49)$$

Where y_i is the unweighted non-heating end-use energy of household *i* in the fuel sub-sample of the 1996 English House Condition Survey, β_0 is the intercept, in this case the group mean for electricity use in households without electric heating (3394 kwh/year), and e_i is the residual for household *i.* To proceed, the researcher assumes that the residuals in the dataset add up to zero and are normally distributed. This assumption can be visualised as the sum of the distances between actual household electricity values *y* above and below the mean β_0:

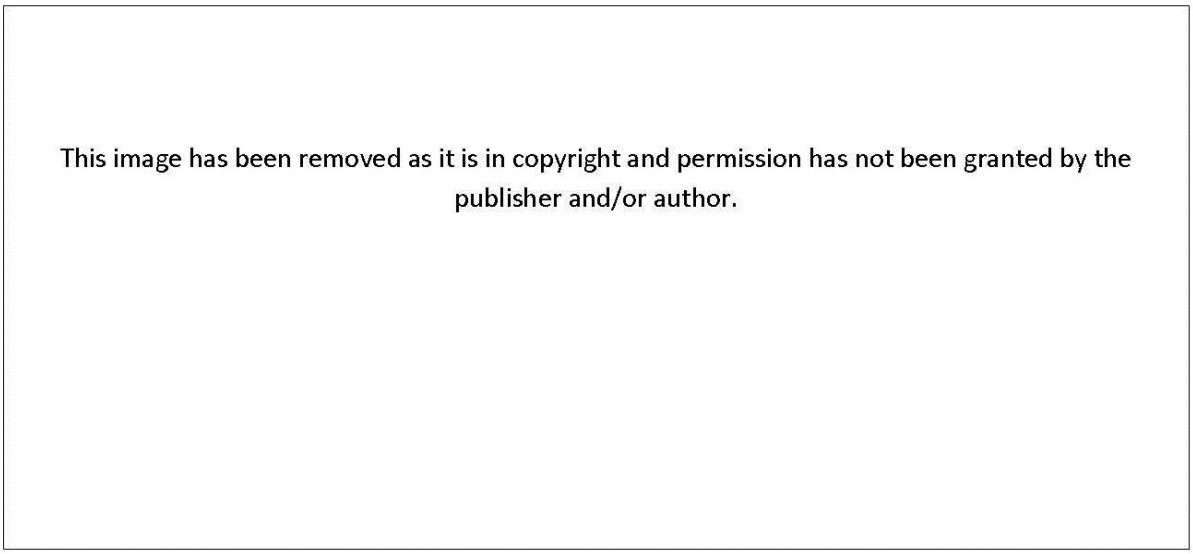


Figure 5.2: Group mean and residuals (Steele, 2011)

Transforming this "empty model" into a multilevel structure using area classifications is straightforward. Each individual household belongs to a unique LLSOA that has been assigned an area classification by the Office for National Statistics from characteristics aggregated from the 2001 United Kingdom Census. Each individual is assign to a group *j.* The residual is also split between the two levels in the new data structure. Area-level residuals are called group random effects u_j and individual residuals within groups are designated as e_{ij}. The equation above turns into:

$$y_{ij} = \beta_0 + u_j + e_{ij} \quad (50)$$

Instead of one single mean β_0 for all non-heating end-use energy *y*, there are several means. The mean of *y* for each group *j* is $\beta_0 + u_j$. Therefore, the group random effect u_j represents the difference between the mean for group *j* and the overall mean value. The individual-level residual

changes from one that is based just on the overall mean to one that is based on the difference between the household *i* and the mean energy use for the area classification group *j* to which the household uniquely belongs. A basic representation of electricity use that is representative of non-heating end-use energy is below (with electricity use less than 20,000 kilowatt-hours per year):

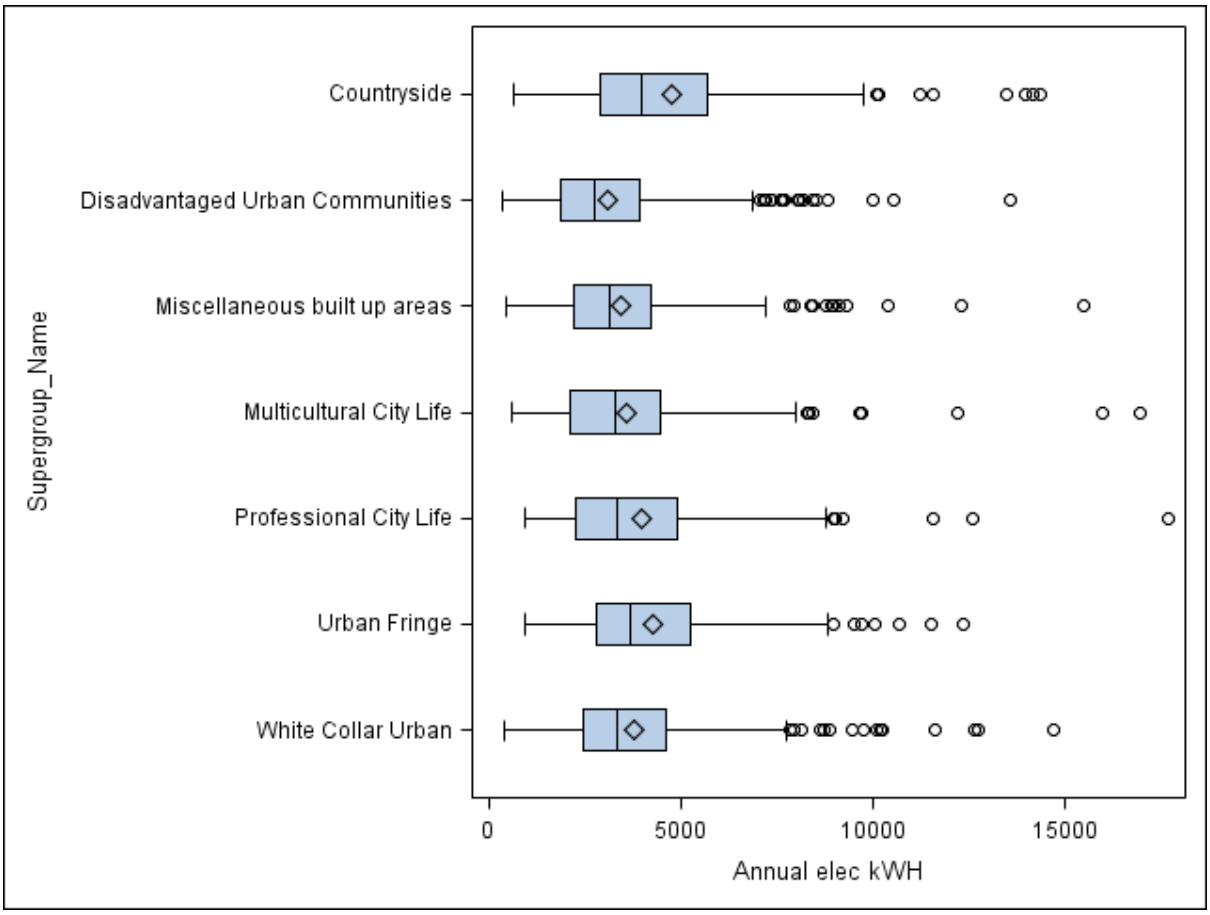


Figure 5.3: Box plots for 2001 ONS Area Classification supergroups of dwellings in the fuel sub-sample of the 1996 English House Condition Survey

The result is visualised below following (Steele, 2011). Figure 5.4 shows two groups, with the individuals in group 1 as grey squares and the individuals in group 2 as black circles. The overall mean β_0 is represented by a solid line, and the group means are represented by dashed lines, with the mean for group 1 below the line and for group 2 above the line. An example of a residual in this two-level model is also given. y_{42} is case number 4 and belongs to group 2, and residual e_{42} is the difference between the actual value of y_{42} and the mean for group 2.

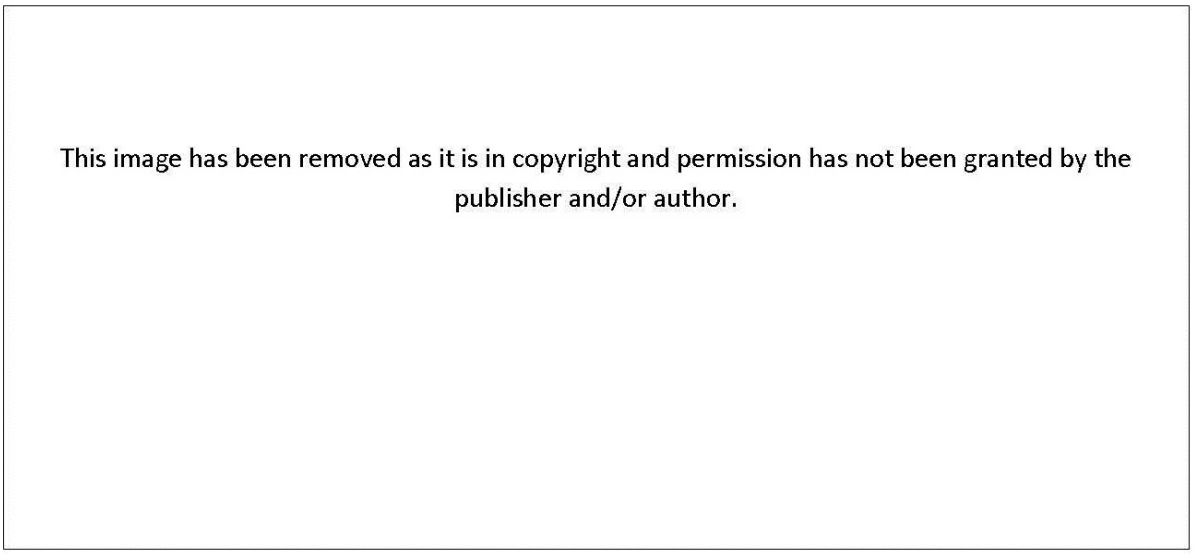


Figure 5.4: Visualisation of group means in a multilevel model (Steele, 2011)

Again, both sets of residuals are assumed to pass a modified version of the parametric test – the individual residuals e_{ij} and the group random effects u_j are assumed to have a normal distribution with a mean of zero. The total variance of a multilevel model is therefore divided between the within-group variance σ_e^2 and the between-group variance σ_u^2. From these variances, it is possible to say how much correlation there is between individuals that belong to the same group, which is called intra-class correlation.

In Figure 5.4 above, individual residuals of the dependent variable in this research, non-heating end-use energy, within each group are unlikely to have a normal distribution. As in the growth model of lights and appliances developed in SAP2009, a single term that represents the interaction of two independent variables may be the likely outcome of this research. Modelling energy use using one interaction term to predict energy use when this data has been transformed to pass the parametric tests will retain the units (e.g. kilowatt-hours) when they are back-transformed. However, modelling of energy use using more than one predictor variable would result in unit-less regression coefficients, and would not be useful for the purposes of this exercise.

An alternative that is possible within multilevel modelling is to allow for within-group variance or heteroskedastic datasets. As shown in Figure 5.4 above, all the supergroups display a similar "long tail" of non-heating end-use energy consumption. Instead of transforming the dependent and independent variables to achieve a normal distribution, the within-group variance can instead depend on one or more independent variables. There are several candidates for this type of variable that itself predicts additional variance in energy use around a regression equation. For example, one theory that can be explored is the influence of building age – as the dwelling is further removed from

its original operational parameters and target socioeconomic market, the variance in the way that it is used and who occupies is greater, and therefore non-heating end-use energy will vary more widely as well.

5.5.5.3 Adding group-level and individual-level independent variables

The nature of variability at the individual level and at the group level was discussed earlier as the foundation of multilevel modelling, but in practice these models use independent variables at the individual level, at the group level, or both. One can add in only individual level variables and control for group effects. The researcher can also go further and add in group level variables, either for a different or the same measure as the individual level variable, to measure group factors that are causing an effect at the same time. Using variables at both levels allow for exploration of contextual effects for households grouped into areas.

One can start to add individual level variables - for example usable floor area – to the "empty" model described in 5.5.5.2 to predict energy use y_{ij} as:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_j + e_{ij} \quad (51)$$

where x_{1ij} is the independent variable of total floor area of all individual households *i* in groups *j* with β_1 as the within-group effect of x_1. This approach that includes only a variable at the individual level can control for clustering or area classification. It can also quantify the effect of belonging to a classification and to query if the effect of household size, for instance, varies across classifications. However, there may be other forces that are "unmeasurable" and "unknowable" at play that this approach cannot take into account.

When group-level variables are introduced to a multilevel model, this problem can begin to be addressed. A second variable, for example, population density, that belongs to the group, can be added as a group level variable as:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j + e_{ij} \quad (52)$$

where x_{2j} is the independent variable of population density in group *j* with β_2 as the contextual effect of x_2. With this structure, the effect of this group-level variable can be explored while allowing for the possibility that energy use may also be influenced by unmeasured group factors. Variables defined at the group level are often called contextual variables and their effects on the outcome – energy consumption in this case - are called contextual effects.

This thesis has previously outlined how the variables that it wishes to explore are those that are shared between datasets of individual household energy consumption and aggregated consumption figures that share similar independent variables. If, for example, the model uses the number of habitable rooms instead of total floor area, these values are available in both the 1996 EHCS and the 2001 Census and therefore, the terms x_{1ij} and x_{2j} are of the same individual level variable and group mean variable, written as x_{ij} and $\bar{x}_j$:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 \bar{x}_j + u_j + e_{ij} \quad (53)$$

where β_1 In this model is the within-group effect of *x*, $\beta_2$is the contextual effect of *x* but $\beta_1 + \beta_2$ is the between-group effect of *x*. By re-arranging the formula to:

$$y_{ij} = \beta_0 + \beta_1 (x_{ij} - \bar{x}_j) + (\beta_1 + \beta_2)\bar{x}_j + u_j + e_{ij} \quad (54)$$

it is possible to measure the between-group effects $\beta_1 + \beta_2$ of household size in a way that was not possible in a linear regression model or in a multilevel model with different independent variables. However, the model is now again limited to one term to represent more than one independent variable such as *TFAxN* used in growth modelling in the previous section.

5.5.5.4 Applying multilevel methods to the annual and decennial options

Using multilevel models can be appropriate in both the annual and the decennial options. Applying the data to the decennial option is more straightforward. The area classifications are updated once every ten years, and the aggregate data that matches the independent variables in the survey dataset is only collected every ten years as part of the United Kingdom Census. In addition, energy consumption data from housing surveys is collected sporadically, and the 1996 EHCS energy sub-sample collection during the years that led up to the 2001 Census is a good match to the aggregate data.

There are possibilities to create algorithms for an annual option, although this would be more unstable. To create a model relevant for years between census takings, one could use more recent surveys to estimate the energy use of homes in the EHCS in later years and the change in the number of dwellings in each super output area. This would require a constancy assumption – that the dwellings in the housing survey would vary from one another in a similar way in later years, and the addition or reduction of homes in a super output area would occur in the same proportions as the existing area. This would have the consequence of holding the clusters in the area classification constant between their construction from decennial censuses.

The application of multilevel models can be beneficial to test a hypothesis that both within-group variation and between-group variation is occurring in non-heating end-use energy. The main benefit of the technique is that non-normal energy consumption and housing data can be used in a sensitive model by linking variance to an independent variable. A drawback is that in order to estimate the effect of group membership directly, one term that represents more than one independent variable present in both the housing survey dataset and the aggregate dataset is necessary. Therefore, the practice of using the interaction term *TFAxN* is likely to continue in a multilevel model. Nevertheless, the multilevel model should be strongly favoured as a useful approach for estimating non-heating end-use energy in households.

5.6 Advancing to detailed analysis

In this chapter, the broad model options and methodological options for using individual-level and area-level data were presented. An annually-updated option and a decennially -updated every ten years - option were presented as the broad model choices. The methodological options considered were econometric, technological, archetypal, ecological, linear regression, and multilevel modelling. Econometric and technological methods were discarded, and the other four methodological options were evaluated against their ability to fulfil the two broad model options using the available data in England after the dissemination of data from the 2001 Census. Finally, the ability of the methodological options to use individual household data and aggregate data separately and simultaneously was considered, as were any transformations necessary in order to make the data approximately parametric for regression analysis.

The following chapter implements the four methodological options remaining and for each composes an algorithm to predict single-household non-heating end-use energy. After this process is complete, a critical assessment is made of the four options, and of whether the extra variability predicted by new variables or by classifying groups is real or due to chance. For example, the linear and polynomial regression and archetypal methods used by BREDEM-8/12 and BREHOMES are nested because the archetypal method adds an extra variable, the scaling factor, to the linear model. This will in turn lead to a schedule for further work in the field for the use of both individual-level and area-level explanatory variables for predicting domestic non-heating energy consumption and its implications for data collection, data protection, privacy, and verification in the future.

Chapter 6 - Dataset validity and preparation

6.1 Introduction

The previous chapters explained how the related work and literature on the estimation of household energy consumption in single households as the building blocks of a bottom-up stock model of non-heating end-use energy led to a hypothesis that area typologies have an effect on consumption beyond household size. They also detailed the data available to the researcher from housing surveys, expenditure surveys, aggregate consumption and census statistics, and area classification systems. Finally, a series of methods was investigated for using either individual household or area level data, or both for exploration in decennially revised model options and annually revised model options. Four were selected for detailed investigation in this chapter.

Chapter 5 also outlined the options available for independent variables that predict domestic non-heating end-use energy, and told the history of using household size as a predictor of appliances, lighting, electronics, and cooking end-uses. The interaction term in use by BREDEM and SAP that combines the usable floor area and the numbers of occupants (themselves modelled using the usable floor area) was investigated in relation to the statistical models considered. The interaction term has been extremely useful in the development of algorithms that measured domestic non-heating end-use energy since the 1980s (Henderson and Shorrock, 1986b, Shorrock and Anderson, 1995, Anderson et al., 1996, BRE, 2010) that combined these two independent variables. This had benefits for non-linear regression to reduce the effect of household size on energy consumption as households become very large, and for measuring between-group effects in multilevel modelling.

This chapter will show how the data available can be adapted for use in the four methods previously explored in this thesis (archetypal, ecological, single-level regression, and multilevel-level regression). This will enable the interaction term (*size of household measured by the built form x size of household measured by occupation)* to be retained in the comparison of the four methods. All the previous models of domestic non-heating end-use energy modelled the number of occupants in a regression model dependent on the total floor area. These models were created in the context of buildings in England and exclusively for the estimation of newly constructed and yet-to-be occupied buildings. The model for the energy rating of existing buildings, called reduced SAP, does not estimate non-heating end-use energy consumption and instead uses metered consumption to

estimate its effect on heating demand via internal gains (BRE, 2010). However, the estimation of scaling factors through area-level variables and area classification depends more on existing buildings, as the housing stock only grows by around one-half of one percent per year (Department for Communities and Local Government, 2011a). Therefore, the interaction term in this thesis will use the total number of occupants as measured in housing surveys and aggregate statistics instead of modelling the number of occupants based on dwelling size.

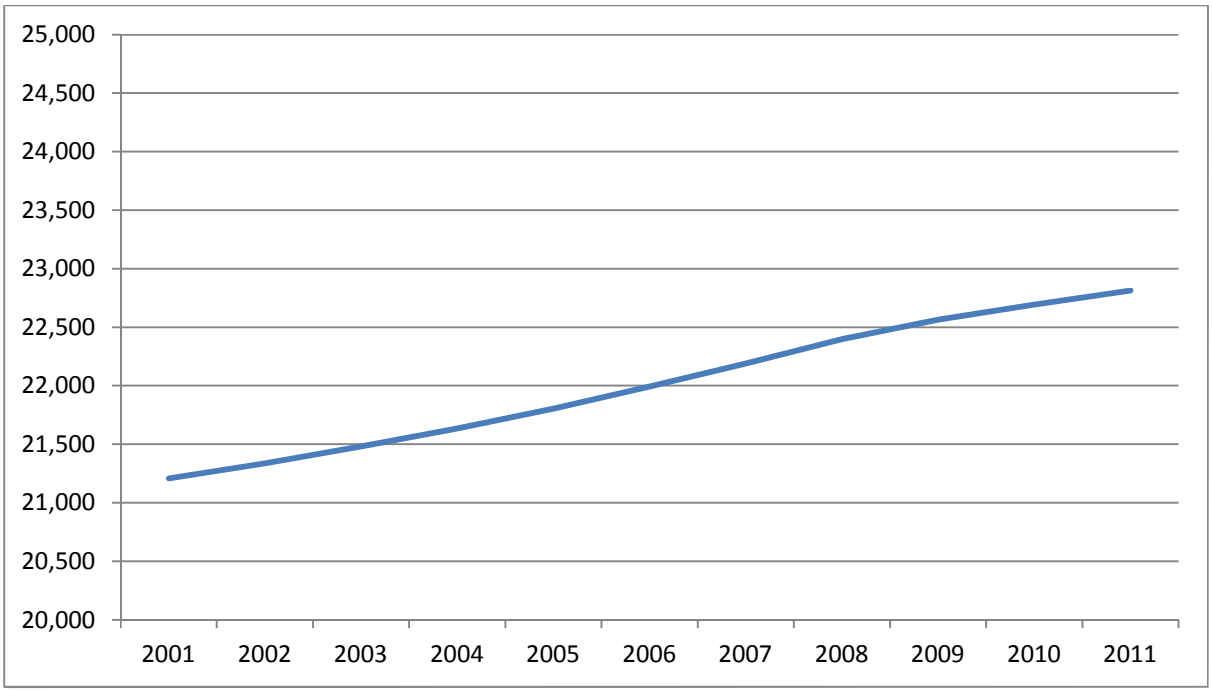


Figure 6.1: Total housing stock in England (thousands of dwellings) 2001-2011

The individual-level data available from the 1996 English House Condition Survey was created out of a complex design using stratified random sampling and not simple random sampling. If a survey has a complex design, the methods that should be used for analysis are slightly different – instead of assuming a random sampling from a theoretically infinite population, the sampling is deliberately chosen from segments of a defined population – England – and given weights, or grossing factors, to enable analysis to describe that chosen population. However, for the purposes of this study, the dataset is reduced to ensure that only homes that do not use electricity for heating are included, and the population represented is reduced. After examination of sample distributions, the conclusion is reached that the assumption that the complex design describes the entire English population no longer has as much force, and that approaches that assume simple random sampling are also valid ways of approaching the modelling of domestic non-heating end-use energy in England.

Finally, this chapter demonstrates the method for preparing the data for the annual option described in previous chapters. This involved the estimation of the amount of non-heating end-use

energy that homes surveyed in 1996 were likely to have used in 2008, using evidence from expenditure surveys taken during 2008. This updating can lead in the future to refinement of models in response to the comparison of a bottom-up housing stock model with aggregated energy use measured in small area datasets available for census areas in England. At the end of this chapter, all of the datasets are prepared for analysis in later chapters.

6.2 Conditions of membership of the datasets

Each of the main datasets that are being considered must reflect the use of electricity for non-heating end-uses. There was a multi-criterion filtering analysis put onto the data in order to reduce confounding factors. If there is a strong probability that electricity in an individual-level housing survey is being used for heating end-uses, that case will be excluded from the analysis. Likewise, if there are many homes that report electricity being used as a heating fuel inside of a small area that reports total electricity use in the residential sector, then that area will not be considered.

The 1996 English House Condition Survey (Department of the Environment Transport and the Regions, 2000b) developed a stratified sampling method based around the strata of age (pre- or post-1945), tenure (private or public), dwelling type (house or flat) and government office region (8 regions). More common combinations of the these characteristics were undersampled, and conversely less common combinations were oversampled. In all, 30,433 addresses were identified for possible inclusion in the housing survey, of which a core sample was developed of 12,131. The fuel sample consists of 3,676 homes that were selected for the fuel survey. Annual averages for fuel use for electricity and natural gas were calculated using up to 9 consecutive quarters of metered fuel data. The 2,531 homes with the most reliable data were given gross weighting factors in proportion to the individual homes' representativeness in the population.

In order to reduce the dataset down to just those homes that use electricity for non-heating end-use energy, 1,277 of the 3,676 cases were removed. The first group to be removed were those without any electricity data: missing values and zero values for reported annual electricity usage. The second group to be removed were those homes that reported using electric heaters or storage for central heating or as a secondary or "top-up" heating source. The third group to be removed were those that either did not have central heating or did not know the fuel source for their home, or did not report any use of natural gas, and it was possible that electricity could be used as a heating fuel. There are other heating fuels that are used besides natural gas and electricity, and these were more common in 1996 than at the present day, but it would be impossible to determine which fuel might be used, and therefore all of these cases were deleted. Therefore the dataset considered consisted

only of homes with natural gas central heating and no reported ancillary electric heating system. This reduced the number of cases down to 2,399 of which 1,776 had gross weighting factors attached to them.

The 2008 Living Costs and Food Survey (Office for National Statistics and Department for Environment Food and Rural Affairs, 2010) interviewed 5,843 households and asked them about the last expenditure they made on electricity. Those who reported using electricity as a heating fuel were excluded, as were those who reported paying for electricity by slot meter or prepayment card. The latter cases were excluded as the payments could not be attached to time periods of electricity use such as months or quarters. This reduced the number of households to 4,929. No grossing factors or weights were created in conjunction with this survey.

The Department for Energy and Climate Change has produced aggregate data for domestic energy use since 2005 (Department for Energy and Climate Change, 2010b). These figures are produced for all 34,378 Lower Layer Super Output Areas (LLSOAs) in England. An LLSOA with around 3,000 households created as a geographic unit by the Office of National Statistics for use with the 2001 Census. The 2001 census contains data on the presence of central heating in homes. If there was a small but significant amount of people who reported not having central heating, then that census area was considered likely to have a significant uplift on the amount of homes that use electricity for heating end-uses. 31% of homes in 2001 without central heating used electricity as the heating fuel, compared to 10% of homes with central heating. By 2006, the proportion of homes without central heating had decreased to 5% and continues to diminish (Shorrock and Utley, 2008). Therefore, it was assumed that areas with more than 95% central heating in 2001 would have the total number homes using electricity as a heating fuel by an acceptably low amount (a maximum of 2.2%). This reduced the number of LLSOAs in England considered to 10,350. At 99%, this number would reduce even further to 1,775 areas and to a maximum of 0.8% of all homes using electricity as their heating fuel. These two options required testing to see if they were reasonably representative of all LLSOAs.

These remaining areas at the 95% and 99% confidence levels are tested if there are at least five areas that are in each 2001 Area Classification Supergroup and Region. It was not expected that all of the various combinations are satisfied (e.g. it is expected that there will be no LLSOAs defined as “Group 1 – Countryside” in the London region). This expectation can be shown in the cross-tabulation of homes that had location data attached to them in the 1996 English House Condition Survey. Only those homes that were taken into the core sample of homes that had a full physical survey, a full interview, and a local housing market assessment had this data attached to them.

Those homes that only took part in the fuel survey and were not part of the core sample did not have the data attached (McIntyre, 2011).

In the survey data in the 1996 English House Condition Survey, the 2001 Area Classifications and the Government Office Regions are well represented in the homes that only use electricity for non-heating end-use energy. From an examination of the data, there was an expected lack of Countryside areas in London and Merseyside and well as the lack of Multicultural City Life in the South West and surprisingly in White Collar Urban areas in London. However, given the total number of census areas classified as White Collar Urban in London in Table 6.1 below, the low number of cases is not surprising. This appears to be an anomaly, as every other region has around the expected number of cases for that supergroup. However, there is not any scope to impute data from other regions into the London region because the nature of a world city like London should be expected to be different from other towns and cities in the rest of England.

Table 6.1: Cases in the 1996 English House Condition Survey cross-tabulated across the 2001 ONS LLSOA Area Classification Supergroup and Government Office Region for England as defined in 1996 (North West and Merseyside were merged, and Eastern renamed East of England in 2001)

Government Office Region	2001 ONS LLSOA Area Classification Supergroup							
Values Frequency Expected Percent	Countryside	Disadvantaged Urban Communities	Miscellaneous built up areas	Multicultural City Life	Professional City Life	Urban Fringe	White Collar Urban	Total
North East	16	146	44	4	7	19	51	287
	26.22	63.879	64.284	37.457	24.094	25.106	45.96	
	0.56	5.15	1.55	0.14	0.25	0.67	1.80	10.12
Yorkshire and the Humber	23	97	75	34	18	22	56	325
	29.691	72.337	72.795	42.416	27.284	28.43	52.046	
	0.81	3.42	2.65	1.20	0.63	0.78	1.98	11.46
North West	14	79	77	33	6	18	54	281
	25.672	62.544	62.94	36.674	23.59	24.581	45	
	0.49	2.79	2.72	1.16	0.21	0.63	1.90	9.91
East Midlands	41	78	59	34	7	31	52	302
	27.59	67.218	67.644	39.414	25.353	26.418	48.363	
	1.45	2.75	2.08	1.20	0.25	1.09	1.83	10.65

West Midlands	28	70	75	55	15	31	65	339
	30.97	75.453	75.931	44.243	28.459	29.655	54.288	
	0.99	2.47	2.65	1.94	0.53	1.09	2.29	11.96
South West	52	32	73	3	14	19	61	254
	23.205	56.534	56.892	33.15	21.323	22.219	40.676	
	1.83	1.13	2.57	0.11	0.49	0.67	2.15	8.96
Eastern	54	28	86	10	15	36	45	274
	25.032	60.986	61.372	35.76	23.002	23.969	43.879	
	1.90	0.99	3.03	0.35	0.53	1.27	1.59	9.66
South East	30	28	73	4	12	30	44	221
	20.19	49.189	49.501	28.843	18.553	19.333	35.391	
	1.06	0.99	2.57	0.14	0.42	1.06	1.55	7.80
London	0	9	19	188	136	28	1	381
	34.807	84.801	85.339	49.725	31.985	33.329	61.014	
	0.00	0.32	0.67	6.63	4.80	0.99	0.04	13.44
Merseyside	1	64	54	5	8	14	25	171
	15.622	38.06	38.302	22.317	14.356	14.959	27.384	
	0.04	2.26	1.90	0.18	0.28	0.49	0.88	6.03
Total	259	631	635	370	238	248	454	2835
	9.14	22.26	22.40	13.05	8.40	8.75	16.01	100.00

To examine if the rows and columns are independent, the expected counts are computed under the null hypothesis. The expected frequency was computed using the PROC FREQ operation in SAS (SAS Institute, 2011) under the null hypothesis that the row and column variables are independent using the formula $\left\{e_{ij} = \frac{n_i n_j}{n}\right\}$ (equation 55) for sample size *n*, the total number of cases in row n_i and the total number of cases in columns n_j. The chi-square statistic is calculated as

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} \qquad (56)$$

Where n_{ij} is the observed number of cases in the table cell (*i,j*). If the rows and columns are independent from each other, then the chi-square statistic will be large, and then the statistic can be compared to a probability table for degrees of freedom (*R*-1)(*C*-1) for the total number of rows *R* and the total number of columns *C.* For the cases with location data in the 1996 English House Condition

Survey, $\chi^2 = 1471$ for DF(54) with a p-value < 0.01. Therefore the null hypothesis, that supergroup membership and regional membership are correlated, is rejected in favour of the alternative hypothesis that the two are independent from one another.

The three possibilities for including census areas that have aggregate electricity consumption data were assessed, and the most restrictive option requiring 99 percent of the census area to have central heating resulted in not enough areas throughout the cross-tabulation. There were several supergroups, especially in the Miscellaneous Built Up Areas and Multicultural City Life supergroups, that did not have any areas included. At the next level examined (at least 95 percent must have central heating) there were at least 4 areas in each cross-tabulation that also were covered by the individual household-level dataset, the 1996 EHCS.

Table 6.2: LLSOAs that in the 2001 Census reported more than 95 percent of households having central heating cross-tabulated across 2001 ONS LLSOA Area Classification Supergroup and Government Office Region

Government Office Region	Supergroup Name							
Values Frequency Percent	Countryside	Disadvantaged Urban Communities	Miscellaneous built up areas	Multicultural City Life	Professional City Life	Urban Fringe	White Collar Urban	Total
North East	40	551	74	4	18	209	345	1241
	0.29	3.99	0.54	0.03	0.13	1.51	2.50	8.98
North West	87	204	45	28	17	608	252	1241
	0.63	1.48	0.33	0.20	0.12	4.40	1.82	8.98
Yorkshire and The Humber	150	239	23	4	16	271	304	1007
	1.09	1.73	0.17	0.03	0.12	1.96	2.20	7.29
East Midlands	298	208	59	37	11	433	467	1513
	2.16	1.51	0.43	0.27	0.08	3.13	3.38	10.95
West Midlands	158	148	60	7	7	424	286	1090
	1.14	1.07	0.43	0.05	0.05	3.07	2.07	7.89
East of England	393	190	346	47	45	701	531	2253
	2.84	1.37	2.50	0.34	0.33	5.07	3.84	16.30
London	1	15	91	582	319	391	35	1434
	0.01	0.11	0.66	4.21	2.31	2.83	0.25	10.38
South East	393	134	391	52	57	1342	694	3063
	2.84	0.97	2.83	0.38	0.41	9.71	5.02	22.16
South West	172	34	76	1	12	350	333	978
	1.24	0.25	0.55	0.01	0.09	2.53	2.41	7.08
Total	1692	1723	1165	762	502	4729	3247	13820
	12.24	12.47	8.43	5.51	3.63	34.22	23.49	100.00

6.3 Measure of physical household size

The options for physical household size differ depending on the datasets used. The only dataset used in the creation of the current model that estimates the baseline electricity consumption for non-heating end-uses is the energy sub-sample of the 1996 English House Condition Survey (Henderson, 2009). Out of the variables available for the representation of household size, the number of square metres has the highest resolution. However, most other datasets, including UK census and expenditure surveys such as the Living Costs and Fuel Survey, count the number of habitable rooms.

The assumption was tested that the higher resolution variable leads to a better correlation with non-heating end-use energy. There were two different correlation statistics considered to test the correlation between non-heating end-use energy and the two proposed interactions physical household size, rooms and occupants. The first is the Pearson Product-Moment Correlation, a parametric measure of association for two variables. If it is a parametric measure, then both variables must pass the parametric tests. The second is the Spearman correlation. The Spearman correlation coefficient can be defined as the Pearson correlation coefficient between the ranked variables. This is a correlation test that is independent of the distribution of the two variables considered (Myers et al., 2010, Conover, 1999).

Both correlation procedures measure the strength and the direction of a linear relationship. If one variable X is an exact linear function of another variable Y, a positive relationship exists if the correlation is 1 and a negative relationship exists if the correlation is -1. If there is no linear predictability between the two variables, the correlation is 0.

A histogram of each variable considered shows that the data is non-parametric for all three variables considered, therefore the Spearman Correlation Coefficient should be used to make the comparison. This was generated using the PROC UNIVARIATE procedure in SAS (Nguyen, 2007). One example is found in Figure 6.2 below:

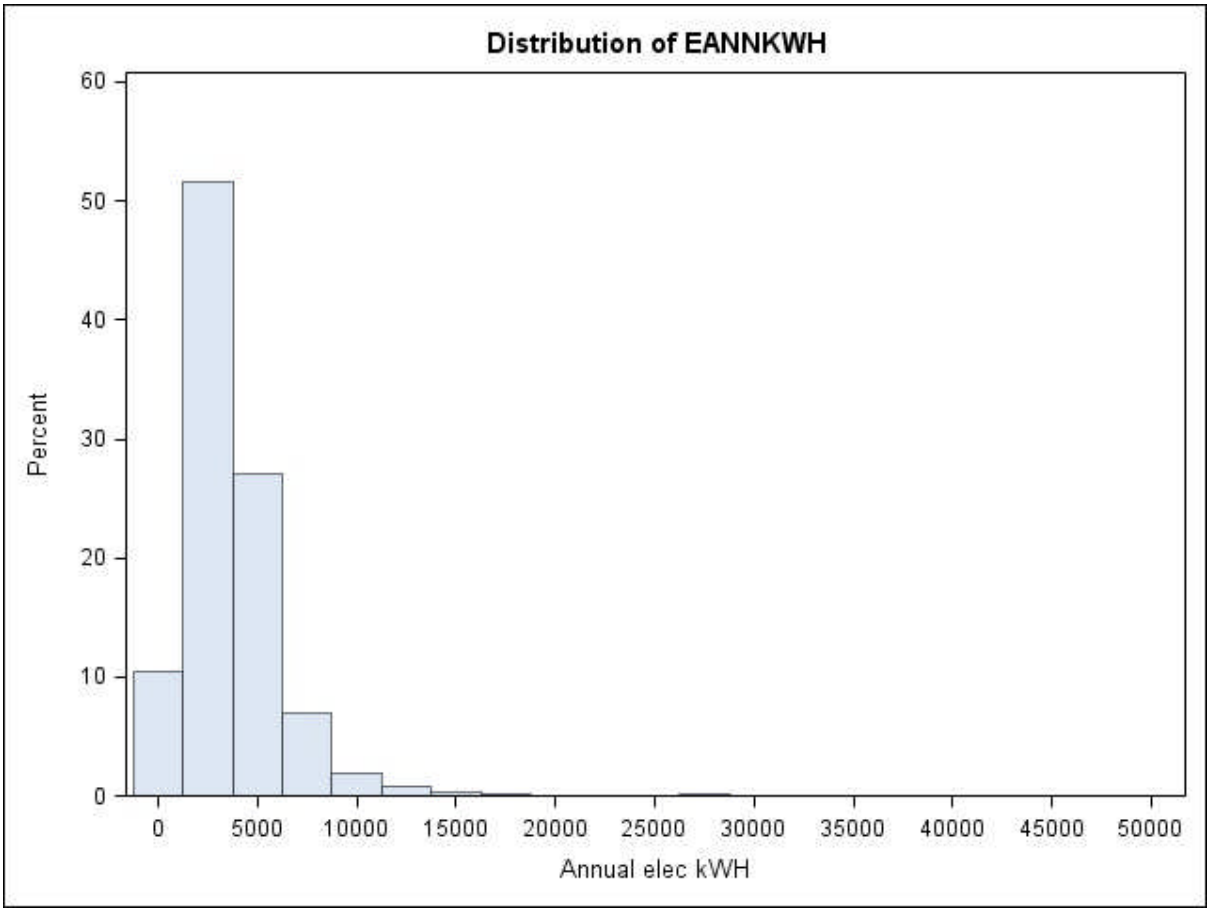


Figure 6.2: Distribution of annual electricity consumption in the fuel sub-sample of the 1996 EHCS

The Spearman Correlation Coefficient ρ is the covariance, or the measurement of how much two variables move together, of the two variables *x* and *y* that are the ordinal rankings of the raw scores X and *Y* over the distribution of all individual cases *i* divided by the product of their standard deviations:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (57)$$

Table 6.3: Spearman correlations of energy and household size from the 1996 EHCS

Variable1	**Variable 2**	ρ	**p-value**
Non-heating end-use energy	Usable floor area	0.37	<0.001
Non-heating end-use energy	Rooms	0.43	<0.001

where non-heating end-use energy is the electricity use of homes that do not report electricity use as their central heating fuel and do not report not having a central heating system.

Both correlations of energy use with physical household size were found to be statistically significant. Therefore, there is no discernible advantage predicted in using floor area over the number of rooms as the measurement of household size in the interaction term (*occupants x size)*.

This thesis will use the number of rooms as the measurement of physical household size instead of usable floor space.

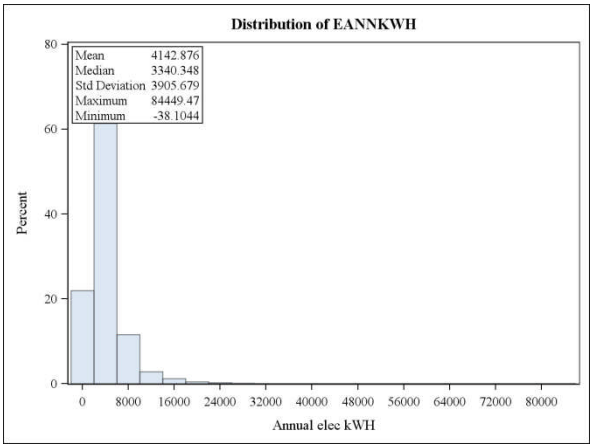
6.4 Analysis of unweighted data and data with grossing factors from the 1996 English House Condition Survey

The 1996 English House Condition Survey had a complex survey design that included four strata of age, tenure, dwelling type, and region (Department of the Environment Transport and the Regions, 2002), and a strong recommendation to always obtain results by using grossing factors appropriate to the survey – in the case of this thesis, the fuel survey (Department of the Environment Transport and the Regions, 2000b). However, the dataset has been substantially altered by removing a number of cases that probably used electricity for heating end-uses and this assertion needs to be re-tested.

There are many reasons why the use of weighting factors in the EHCS fuel sub-sample was found to be undesirable. The use of regression analysis on the data requires the assumption that the sampling is done at random from an infinite population. However, as the gross weighting factors are intended to help approximate the real population of England through a complex survey design, this assumption may lead to incorrect conclusions. Moreover, cases reporting primary or secondary electric heating need to be eliminated, unbalancing the weights assigned during stratification. The use of weighted variables in multilevel analysis must be based solely on the classification scheme, and as the determination of strata in the EHCS and the area classification clustering techniques are different, this could cause difficulties in analysing the data.

The number of cases with a grossing factor was reduced from 2,531 to 1,776 before dealing with outliers, high leverage points, and transformation; this results in a loss of 27.5 percent of the fuel sample. The population represented by these cases reduces from 19,361,059 households to 14,018,680 households, a reduction of 29.8 percent. This reduction is similar enough to warrant further investigation that the reduction of households is evenly spread across the population. This is done by testing the frequency distribution of the dependent variables of the original and reduced datasets to see if these are similarly distributed.

The four different options were compared: all cases of domestic electricity use, all cases with grossing factors, all cases in the reduced dataset with only homes that use electricity for non-heating end-uses, and all cases in the reduced dataset with grossing factors.



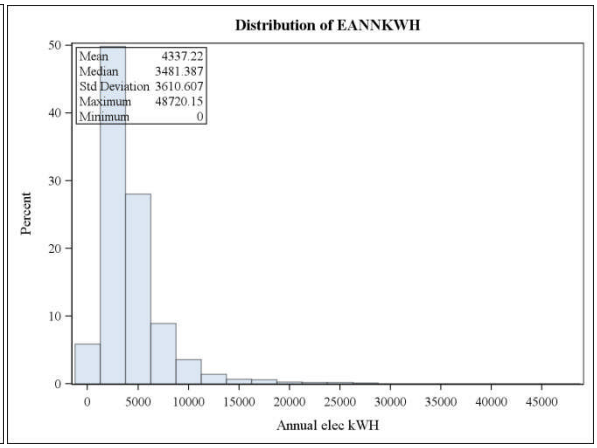
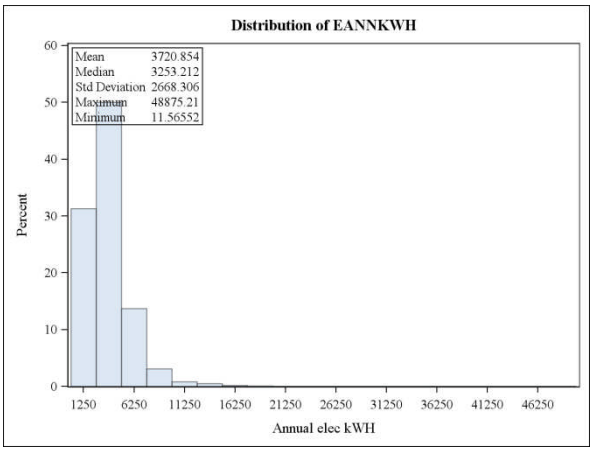


Figure 6.3 (left): Electricity use distribution, all cases

Figure 6.4 (right): Electricity use distribution, all cases that have a grossing factor



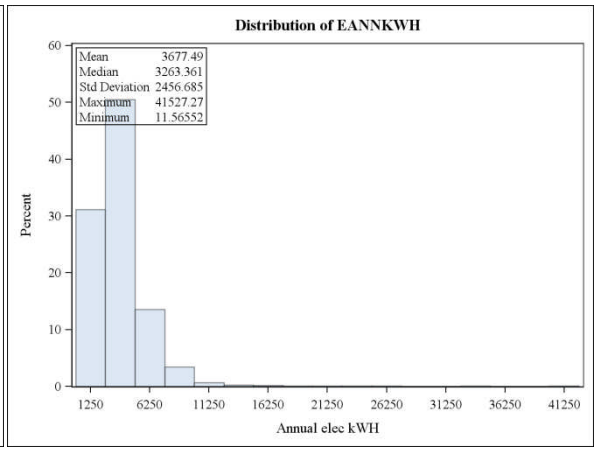


Figure 6.5 (left): Electricity use distribution, reduced set (first reduction), all cases

Figure 6.6 (right): Electricity use distribution, reduced set (first reduction), all cases that have a grossing factor

The skew, means and medians are very similar in the reduced dataset, and therefore one can assume confident that the removed cases have been evenly distributed across the original dataset. However, there are discernible differences in the median and the distribution in the dataset that contains all the cases with outliers and high leverage points included.

There is a similar change in the shape of the distribution and reduction of the median from the full to reduced dataset in both the situations where all cases are included and where only weighted cases are included. This indicates that the reduced dataset with the exclusion of homes that use electricity for heating has the same distribution across the full set of cases and the weighted set of cases.

The use of the full fuel sample with stratification without the removal of cases must include grossing factors for the conclusions to be valid. As discussed above, the removal of cases could invalidate any

conclusions because the grossing factors are based on the number of cases sampled in each strata. However, a user of the reduced fuel sample, because the distributions and medians of the dependent variable are similar to the full sample, can choose to use the unweighted dataset in raw form. This opens up the possibility of using the unweighted dataset in multilevel analysis using area classification, which requires a dependent variable whose variance is normally distributed around a mean of zero.

6.5 Transformations for multilevel regression and parametric data analysis

Parametric tests of housing data are the most reliable way of creating models of electricity use in dwellings. Statistical analysis of housing has the problem of several key variables not passing these parametric tests in raw form and therefore requiring transformation. A parametric test makes assumptions about the population from which the dataset is drawn. In theory, this could be for any type of distribution with known parameters defining a finite or infinite population. In practice, this term is reserved for the parametric tests that are based on the normal distribution of an infinite population, which requires four assumptions to be satisfied (Field and Miles, 2010a). These four assumptions for a normal distribution are:

- Normally distributed data. The majority of scores lie around the centre of distribution. There are two ways that a distribution deviates from normal: first, a lack of symmetrical data, called skew, and too many scores in the tails or middle of the distribution, called kurtosis.
- Homogeneity of variance. The variance of the difference between the expected and actual values of *y* based on the value of independent variable *x* remains constant for any range of *x* chosen.
- Data collected at least at the interval level, if not at the continuous level.
- Independence. The assumption that the behaviour of one individual or case does not influence the other cases.

A histogram of annual electricity use reported in the 1996 EHCS reveals that the data has a strong positive skew, but with a boxplot of annual electricity use by house size, it is clear that some measurement anomalies are present.

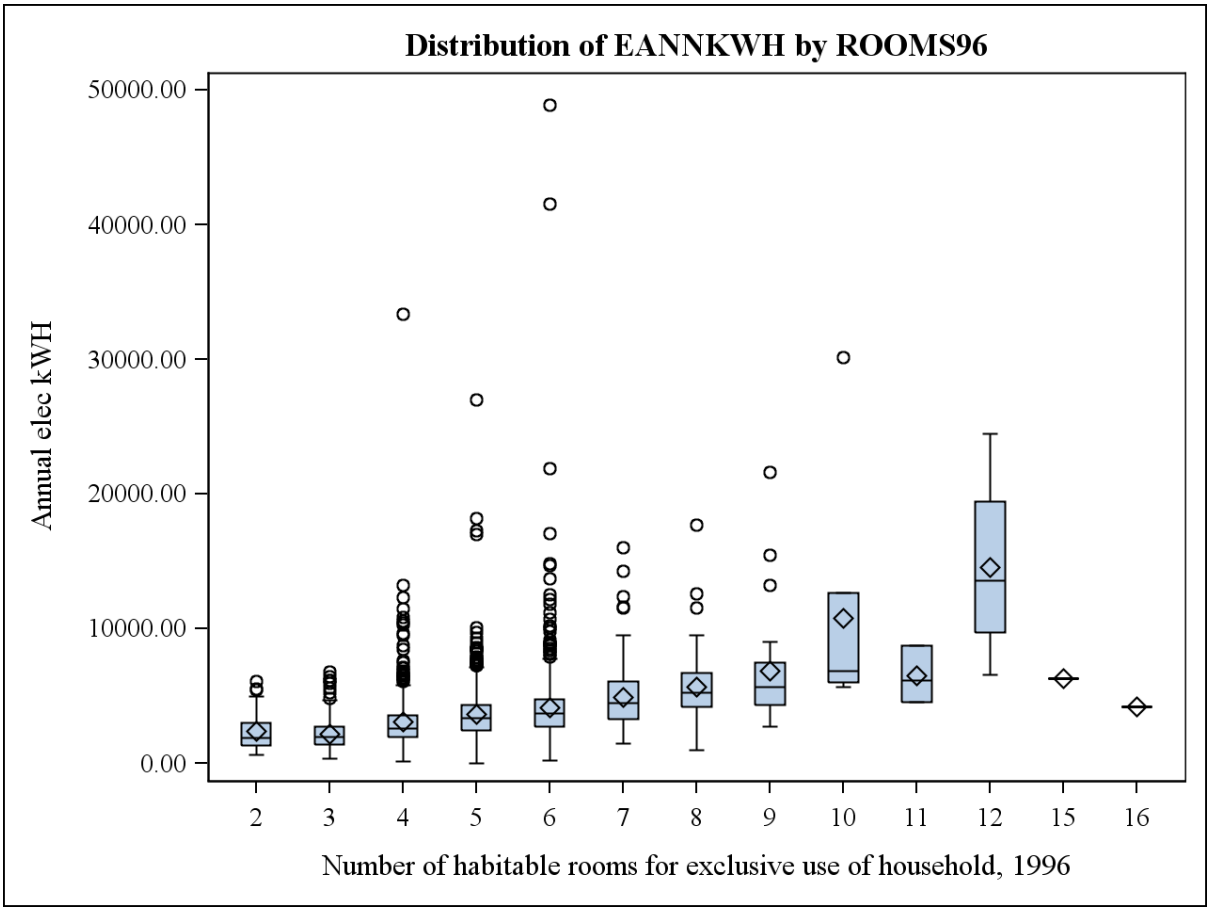


Figure 6.7: Box plot of annual electricity use by number of habitable rooms.

In Figure 6.7, the diamond represents the mean, the line stands for the median, the shaded area is the interquartile range, or the area between the 25th and 75th percentiles, the "whiskers" move to the maximum value below the upper fence, defined as 1.5 interquartile ranges above the 75th percentile, and the minimum value above the lower fence, defined as 1.5 interquartile ranges below the 25th percentile, and dots represented outliers beyond the upper and lower fences for each size group.

6.5.1 Reducing the dataset by excluding outliers and high leverage points

Both outliers and high leverage points are evident from the box plots above. Outliers are instances of the dependent variable, in this case, energy use, that are unlikely to be part of the population whose behaviour is designed to be estimated by a linear or other parametric regression model. Likewise, high leverage points are sizes of homes that are likely not to be part of a population that can be modelled. It was determined that these homes were so unlike the rest of the population that any model could not encompass their characteristics, and therefore they were excluded.

Outliers are clearly evident above the “upper fence” of the box plots, and equally noticeable is the absence of outliers below the “lower fence” of the box plots. However, this variance created by the outliers above the upper fence is constant across all values of household sizes, and the outliers are clustered close to the upper fence, and not dispersed across a large range of values. It is unlikely that the electricity use more than two standard deviations from the mean is limited to electricity use of appliances, lighting, electronics, and cooking in dwelling units. The reasons for this are likely to be:

- Use of electricity for space heating or water heating despite a different heating fuel being reported in the interview survey
- Specialised equipment that operates in a commercial environment, such as computer servers, or in a leisure environment, such as heating of conservatories

Both the dependent variable, non-heating end-use energy, and the interaction term of the two independent variables, the number of rooms, and the numbers of occupants, were assessed for outliers. Both the dependent variable and the interaction term were converted into z-scores. A *z-score* quantifies the original score - in this case, either the dependent variable or the interaction term - in terms of the number of *standard deviations* that that score is from the mean of the distribution. In terms of a formula:

$$z = \frac{x-\mu}{\sigma} \quad (58)$$

Where z is the z-score, x is the raw score, μ is the mean of all cases x, and σ is the standard deviation of x.

Table 6.4: Basic statistics on dependent and independent variables in the fuel sub-sample of the 1996 English House Condition Survey

Variable name	Label name	Num	Mean	Standard Deviation	Minimum	Maximum
EANNKWH	Annual non-heating end-use energy (kWh of electricity, 1996)	2399	3720.85	2668.31	11.5655209	48875.21
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	2370	13.7860759	9.8221100	1.0000000	135.0000000
zeannkwh	Z-score of non-heating end-use energy	2399	-6.94946E-16	1.0000000	-1.3901285	16.9224794
zrooms_hsize96	Z-score of interaction term	2370	-6.399007E-17	1.0000000	-1.3017647	12.3409251
outliere	Number of whole standard deviations outside of the mean energy use	2399	0.0508545	0.3366640	0	3.0000000
leverages	Number of whole standard deviations outside of the mean interaction term	2399	0.0629429	0.3450445	0	3.0000000

If these z-scores are ultimately to be part of a normally distributed set of data with or without transformation, then it can be justifiable to exclude cases in the sample taken by the 1996 English House Condition Survey that are less than 1% likely to be part of the main dataset. According to the curve of a normal distribution, this is defined as more than 2.58 standard deviations from the mean, or a z-score greater than 2.58. For information, outliers in the sample with a value greater than 1 are more than 1.96 standard deviations from the mean and are less than 5% likely to be part of the real population. Outliers in the sample with a value greater than 3 are more than 3.58 standard devations from the mean and are less than one-hundredth of one percent likely to be part of the statistical population.

The exclusion of these cases will only exclude either, or both, the large users of energy (more than 10,600 kilowatt hours compared to a mean of 3,720) or extremely large households (more than a value of 40 for the interaction term compared to a mean of 13.8). The excluded households were more likely to have more people than rooms (37%) than the rest of the sample (1.4%). This indicates

that the outliers are part of a distinct population of households in relation to their use of non-heating end-use energy, and estimations of their energy use should not be a model that estimates the energy use of a single existing or proposed household.

Table 6.5: Outliers of non-heating end-use energy in the fuel sub-sample of the 1996 English House Condition Survey

Number of whole standard deviations outside of the mean energy use				
outlier	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
<1	2335	97.33	2335	97.33
1.00-1.99	26	1.08	2361	98.42
2.00-2.99	18	0.75	2379	99.17
>=3.00	20	0.83	2399	100.00

Table 6.6: Leverage of the interaction term (number of rooms x number of occupants) in the fuel sub-sample of the 1996 English House Condition Survey

Number of whole standard deviations outside of the mean interaction term				
High leverage	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Missing	29	1.21	29	1.21
<1	2278	94.96	2307	96.17
1.00-1.99	46	1.92	2353	98.08
2.00-2.99	33	1.38	2386	99.46
>=3.00	13	0.54	2399	100.00

Table 6.7: Excluded cases from the fuel sub-sample of the 1996 English House Condition Survey

Excluded cases – either household size data missing, or more than 2 standard deviations outside the mean for either the dependent variable or the interaction term				
toexclude	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Not excluded	2293	95.58	2293	95.58
Excluded	106	4.42	2399	100.00

However, the untransformed data, even without these outliers or high leverage interaction terms, remains non-normally distributed. This again is not surprising because of the distribution of house sizes themselves having a positive skew. In order to perform parametric tests on the data such as regression analysis and an analysis of variance, the dependent variable, in this case, electricity use in households, must be approximately normally distributed, have similar amounts of variance throughout the data, be independent from one another, and at least be measured at interval level (Field and Miles, 2010b). A decision was made to transform all of the data in this analysis because transforming only the dependent variable and not the other variables would result in comparisons of geometric means instead of arithmetic means.

It should also be noted that this problem could be mitigated by the inclusion of every data point more than two standard deviations from the mean. The inclusion of these outliers, however, would make energy use models unreliable. Neither solution is satisfactory, but the exclusion of outliers from the dataset is the least bad choice that can be made in this situation. This acknowledges the limitations this data has in predicting energy use in large households or in predicting heavy electricity use in otherwise normally sized households. Clearly, this can be a subject of further targeted sampling using stratifications based on household size and not based on tenure or dwelling types as done in current and past editions of the English House Condition Survey. Later in the thesis, there will be a discursive discussion of the variance between the predicted and measured electricity use both nationally in England and geographically by census area.

6.5.2 Transformation of the dataset to an approximately normal distribution

Even after reducing the dataset, the dependent variable of non-heating end-use energy is not normally distributed upon a further visual inspection of the frequency distribution. As the variables are positively skewed, there were initial approaches chosen for transformation of the data with a positive skew, with a square root transformation proving to have a more normal distribution than a natural logarithm transformation or a fourth root transformation. The square root transformation, after examination of histograms, quantile-quantile (Q-Q) plots, probability plots and the values of skew and kurtosis, produces the closest approximation to normally distributed set of scores for annual electricity use for non-heating end-use energy.

Histograms represent the frequency distribution of total values across a range of pre-determined or automatically generated intervals. A normally distributed histogram should not have too many values to one side of the median value (skew), nor have too many of the mode range of values at the mean (kurtosis). Figure 6.8 below is a histogram showing the distribution of the dependent variable after removal of outliers and high leverage points:

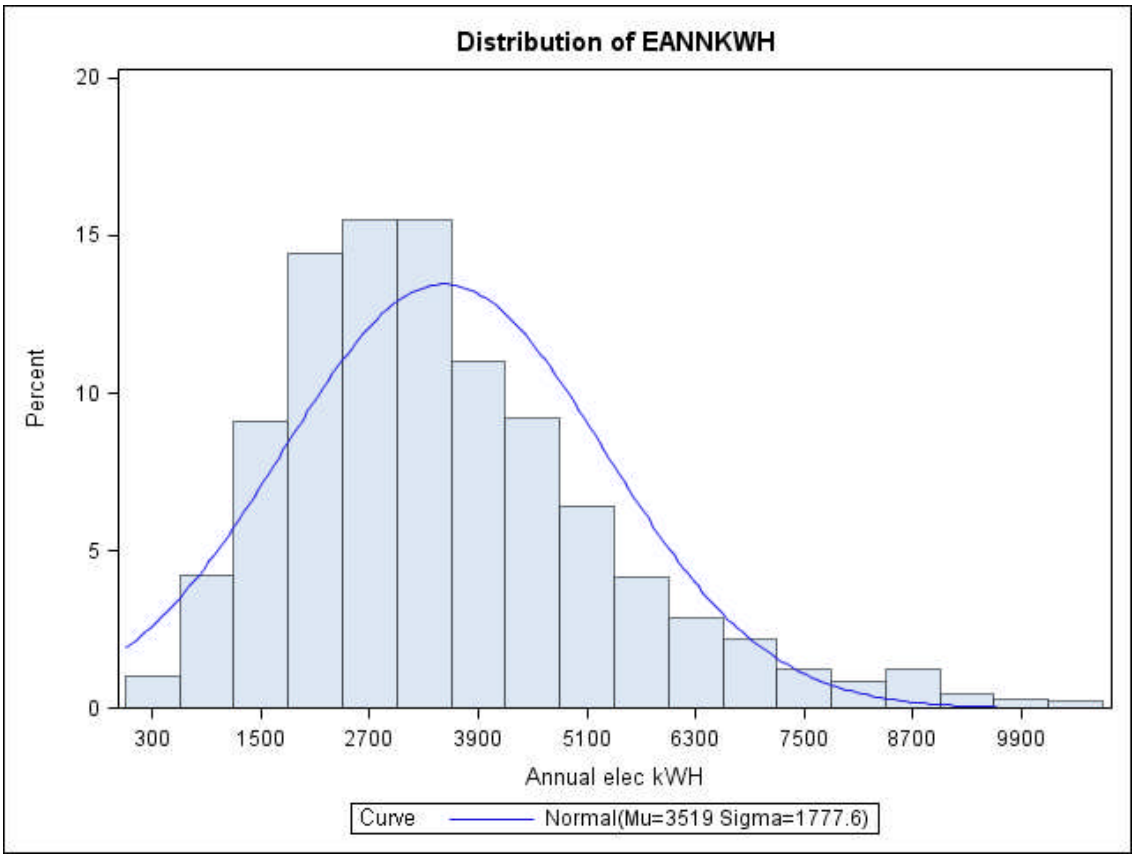


Figure 6.8: Histogram of the dependent variable (electricity in homes that do not use electricity for heating end-uses), untransformed, outliers and high leverage points removed (reduced dataset)

There are also descriptive statistics for the amount of skew and kurtosis in the variable that can be computed directly from the dataset and are defined by standard algorithms calculated by the PROC UNIVARIATE procedure in SAS for central moments (Fisher, 1973, SAS Institute, 2011). A *kth* central moment is formed of a set of values that critically characterise the properties of a probability distribution. The first four moments measure the mean, the variance of the distribution, the lopsidedness of the distribution, and the flatness of the distribution. Each k^{th} central moment is defined by the algorithm

$$u_k = (X - \mu)^k$$

where *X* is the random variable probability distribution, or list of probabilities associated with each of its possible values with a predicted value μ resulting in a estimated value for that particular moment u_k that represents the population as a whole and not just the sample that contains the data.

Therefore, the methods proposed by Fisher and implemented in PROC UNIVARIATE to put values onto these central moments do not simply standardise the central moment and divide by the sample size of nonmissing values *n.* This enables the researcher to measure the variance, skewness, and kurtosis of the population and not just the sample. This is also important because the assumption of a normal distribution is based on the shape of an infinite population represented by an unweighted sample. This is done by dividing by c_k which gets closer and closer to 1/*n* as sample size *n* gets very large (provided *n* > *k*):

$$c_1 = \frac{1}{n} \ , k = 1 \tag{59}$$

$$c_2 = \frac{1}{n-1} \ , k = 2 \tag{60}$$

$$c_3 = \frac{n}{(n-1)(n-2)} \ , k = 2 \tag{61}$$

$$c_4 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \ , k = 3 \tag{62}$$

These measures can also accept the gross weighting variables created by the 1996 English House Condition Survey in order to deal with the way that a stratified random sample approximates a population. This is also important because visualisations such as histograms made by statistical analysis programmes cannot take into account weighted variables. In this case, the variance σ^2 is the sum of the weighted, squared differences between the value of variable *x* and the weighted mean of all the cases multiplied by c_2. This can be represented by:

$$u_2 = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2 \quad (63)$$

where *n* is the number of nonmissing values for a variable, x_i is the *i*th value of the variable, $\bar{x}_w$ is the weighted mean, and w_i is the weight associated with the *i*th value of the variable. This then can be used to compute the weighted standard deviation $\sigma_w = \sqrt{\sigma^2}$. If the sample is unweighted, then all values of variable w_i are equal to 1.

The score for the amount of skew is the estimate of the average of the standardised third moment $\{g1 = u_3 {u_2}^{-3/2}\}$ about the mean. For any sample taken of sample size *n*, such as the sample taken of energy use in housing as part of the 1996 English House Condition Survey, the skew of the population is denoted as :

$$g1 = c_3 \sum_{i=1}^{n} {w_i}^{3/2} \left(\frac{x_i - \bar{x}_w}{\sigma_w}\right)^3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} {w_i}^{3/2} \left(\frac{x_i - \bar{x}_w}{\sigma_w}\right)^3 \quad (64)$$

where *x* is the raw score, μ is the mean of the entire population *x*, and σ is the standard deviation of *x*.

The value of skewness in the distribution of the population is zero if the distribution is exactly normal. As sample size *n* gets larger, the difference between the value of skewness and zero that is non-significant can also get larger.

The procedure calculates kurtosis in much the same manner. Fisher calculated the theoretical value as the average of the fourth standardised moment $\{g2 = u_4 {u_2}^{-2}\}$. For any sample taken of size *n*, the kurtosis of the population is denoted as:

$$g2 = c_4 \sum_{i=1}^{n} {w_i}^{2} \left(\frac{x_i - \bar{x}_w}{\sigma_w}\right)^4 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} {w_i}^{2} \left(\frac{x_i - \bar{x}_w}{\sigma_w}\right)^4 \quad (65)$$

Again, the value of kurtosis in the distribution of the population is zero if the distribution is exactly normal. As sample size *n* gets larger, the difference between the value of kurtosis and zero can also get larger and become significant.

There is a range of different combinations of transformations that may make the dependent variable (non-heating end-use energy) approximately parametric, therefore enabling linear estimation techniques. A Box-Cox transformation can be used to find the most likely nonlinear transformation of a dependent variable (Box and Cox, 1964). This is defined as:

$$y(\lambda) = \frac{y^{\lambda}-1}{\lambda} \text{ if } \lambda \neq 0,\ y(\lambda) = \ln(y), \text{ if } \lambda = 0 \quad (66)$$

Where *y* is the dependent variable and λ (lambda) is the transformation parameter. The most powerful of these transformations are as follows:

- λ = 1.00: no transformation needed; produces results identical to original data
- λ = 0.50: square root transformation
- λ = 0.33: cube root transformation
- λ = 0.25: fourth root transformation
- λ = 0.00: natural log transformation
- λ = -0.50: reciprocal square root transformation
- λ = -1.00: reciprocal (inverse) transformation

The following PROC TRANSREG procedure modelled the possible transformations of the unweighted, untransformed value of non-heating end-use energy to the nearest "power" transformation value in Figure 6.14 below. The procedure chooses the optimal power transformation by using a maximum likelihood criterion measured against the *f-statistic,* or the variance of mean divided by the mean of the variances to the different lambdas, or power parameters described earlier (Draper and Smith, 1981, SAS Institute, 2011):

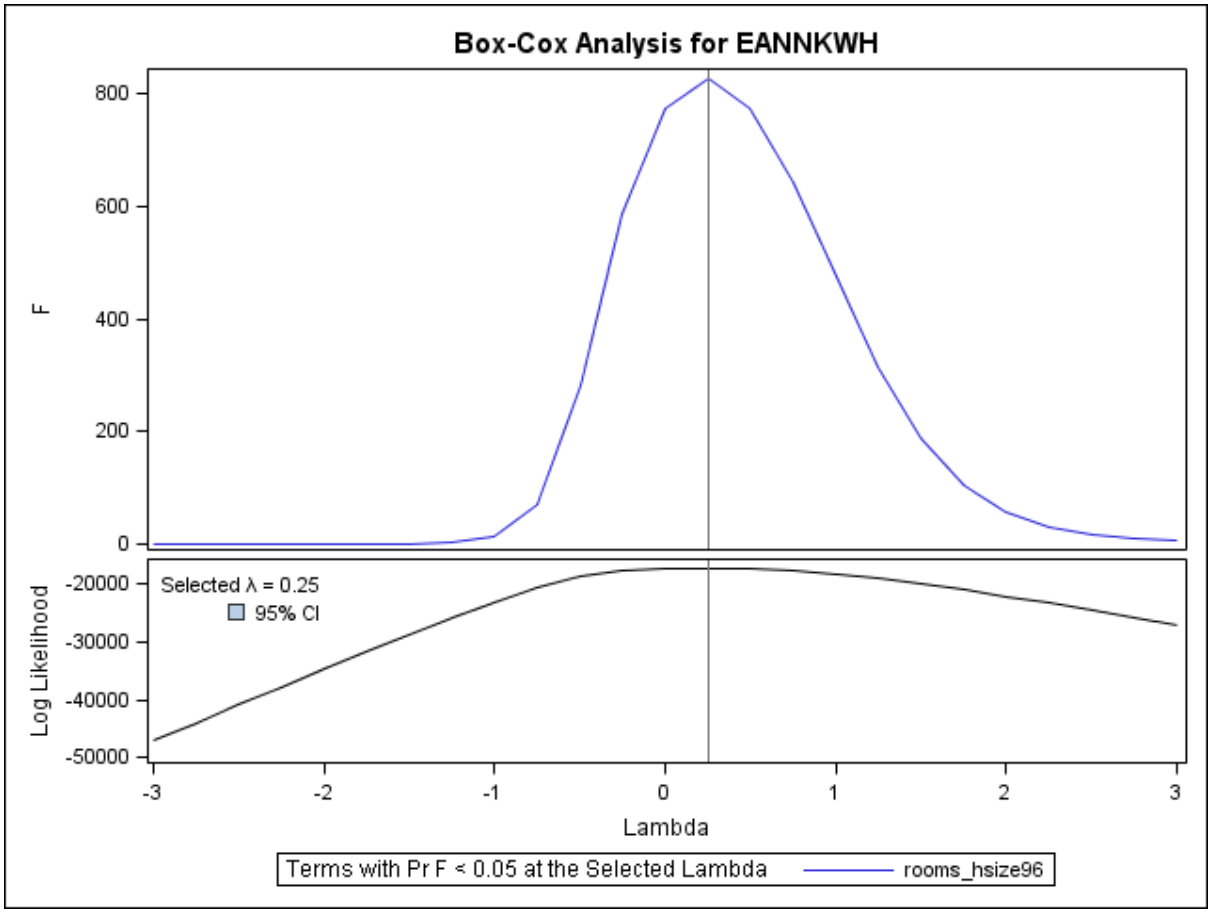


Figure 6.9: Box-Cox transformation of full dataset

The Box-Cox statistics suggest the most likely power transformation for non-heating end-use energy to be the fourth root transformation λ = 0.25. Compared to nearby power parameters, namely the log transformation λ = 0 and the square root transformation λ = 0.5, this indeed is the case when measuring the skewness and kurtosis of the distribution using Fisher's algorithms described above for the unweighted and weighted samples with all points included. However, even using the fourth root, the kurtosis is still far from zero as shown in Table 6.8 below:

Table 6.8: Measurements of skewness and kurtosis in dependent variable, full dataset

Number of cases	Mean	Skewness	Kurtosis	Z-Score: Skewness	Z-Score: Kurtosis	Transformation and Weighting	Outliers and High Leverage Points
2399	3720.85	5.71968	67.2927	114.441	673.487	Unweighted	In
2399	8.05	-0.79211	5.4327	-15.849	54.372	Unweighted, log-transformed	In

Number of cases	Mean	Skewness	Kurtosis	Z-Score: Skewness	Z-Score: Kurtosis	Transformation and Weighting	Outliers and High Leverage Points
2399	58.43	1.53810	8.5455	30.775	85.526	Unweighted, square root-transformed	In
2399	7.56	0.39648	3.0653	7.933	30.678	Unweighted, fourth root-transformed	In
3943.32	6.65958	97.4741	3943.32	114.672	839.682	Weighted	In
8.13	-1.07051	17.0105	8.13	-18.433	146.535	Weighted, log-transformed	In
60.57	2.03704	15.5036	60.57	35.076	133.555	Weighted, square root-transformed	In
7.71	0.73172	7.6399	7.71	12.600	65.813	Weighted, fourth root-transformed	In

Although it is possible to calculate a *z-score* for skew *g1* and kurtosis *g2* $\left\{Z_{g1} = \frac{g1}{SE_{g1}}, Z_{g2} = \frac{g1}{SE_{g2}}\right\}$ (equation 67) , large sample sizes rapidly decrease Fisher's estimate of the standard error of skew and kurtosis $\left\{SE_{g1} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}, SE_{g2} = \sqrt{\frac{2SE_{g1}n^2}{(n-3)(n+5)}}\right\}$ (equation 68), and therefore artificially inflate *z-scores* (Fisher, 1973, Field and Miles, 2010a). In smaller samples, a large *z-score* against the normal distribution curve at the 95% confidence level should be below 2.58. It is assumed that larger sample sizes give a better chance for a normal distribution; that is, as the sample size increases, the skew and kurtosis should decrease. Therefore, the best transformation should simply be the one that minimises the raw score for skewness and kurtosis.

However, the maximum likelihood for the log transformation λ = 0 and the square root transformation λ = 0.5 are very close to the maximum likelihood for the forth root transformation, with only a slightly lower *f-statistic*. This is a notable development: despite the fourth-root transformation being identified as the potentially best transformation, the log-transformation and the square root transformation are much easier transformations in which to form an algorithm, where the untransformed linear equation

$$y = ax + b \tag{69}$$

where *y* is the estimated energy use based on the interaction term *x* with slope *a* and intercept *b* could be used for the growth model $\{y = ax^c\}$ using the log-transformation of *y* and the interaction term as part of the linear model

$$\ln y = \ln a + c \ln x \quad (70)$$

This model is currently used in England to represent non-heating end-use energy, and it does simulate a diminishing rate of increase of energy use in relation to household size as c<1 (BRE, 2010).

In addition, the linear model based on just the square root transformation of the dependent variable and not the interaction term $\{\sqrt{y} = -ax + b\}$ (this sign is expected to be minus to reflect diminishing returns). When back-transformed, this results in the algorithm

$$y = a^2x^2 - 2axb + b^2 \quad (71)$$

However, the fourth root transformation of the dependent variable in a linear regression model $\{\sqrt[4]{y} = -ax + b\}$ would result in an algorithm that would be difficult to interpret as a reliable estimate of human behaviour:

$$y = a^4x^4 - 4a^3x^3b + 6a^2x^2b^2 - 4axb^3 + b^4 \quad (72)$$

Therefore proceeding to a fourth root transformation should be seen as an undesirable outcome of this investigation and should be avoided if necessary.

As the skew and kurtosis are expected to decrease with the exclusion of outliers of the dependent variable and the high leverage points in the interaction term, the fine balance of this term shifts with a re-running of the Box-Cox transformation on the dataset with the outliers and high leverage points excluded:

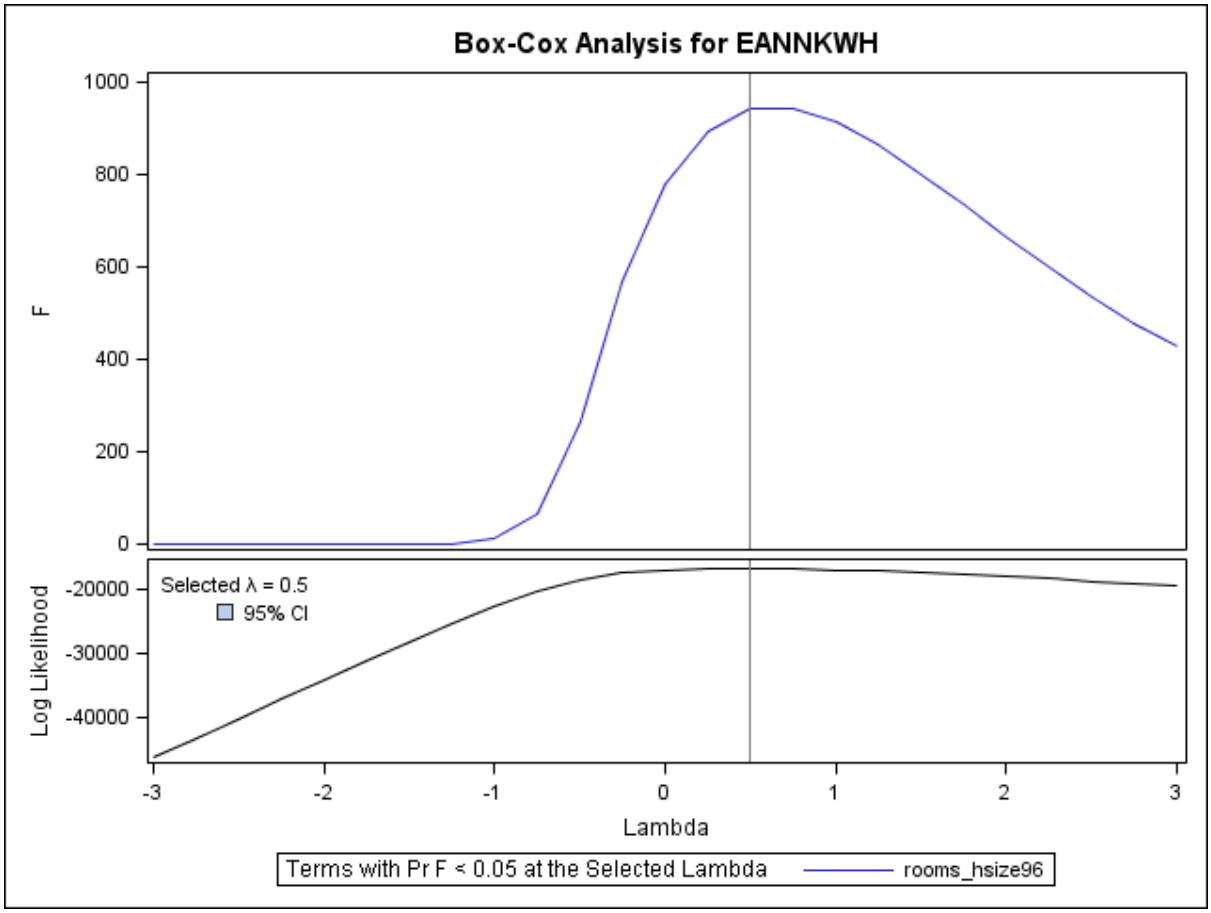


Figure 6.10: Box-Cox transformation of reduced dataset

The calculations of skewness and kurtosis for the new reduced dataset are included in table XXX below, for the dataset, weighted and unweighted. Note that around 25 percent of the cases in this reduced dataset were not given gross household weighting variables:

Table 6.9: Measurements of skewness and kurtosis in dependent variable, reduced dataset

Number of cases	Mean	Skewness	Kurtosis	Z-score: Skewness	Z-score: Kurtosis	Transformation and Weighting	Outliers and High Leverage Points
2293	3467.00	0.98858	1.2087	19.338	11.827	Unweighted	Excluded
2293	8.01	-1.26769	6.2487	-24.798	61.144	Unweighted, log-transformed	Excluded
2293	57.01	0.19234	0.2303	3.763	2.254	Unweighted, square root-transformed	Excluded
2293	7.48	-0.33710	0.9410	-6.594	9.208	Unweighted, fourth root-transformed	Excluded

Number of cases	Mean	Skewness	Kurtosis	Z-score: Skewness	Z-score: Kurtosis	Transformation and Weighting	Outliers and High Leverage Points
1701	3732.72	1.72103	7.9043	29.003	66.642	Weighted	Excluded
1701	8.10	-1.62634	19.8182	-27.408	167.089	Weighted, log-transformed	Excluded
1701	59.39	0.72929	4.4087	12.290	37.170	Weighted, square root-transformed	Excluded
1701	7.65	0.01653	4.9610	0.279	41.826	Weighted, fourth root-transformed	Excluded

The raw scores for skewness and kurtosis are much closer to zero with less than five percent of the dataset removed in reduced form than in the full housing survey that contains the outliers and high leverage points. The best performing transformation and weighting strategy is the square-root transformed, unweighted option. The surprising result is that gross weightings created as part of the stratification procedure increases the deviation from a normal distribution. As the guidance to the English House Condition Survey and subsequent housing surveys that cover England make clear, the application of weights is essential to the understanding of housing when creating a single-level model (Department of the Environment Transport and the Regions, 2002, Department of the Environment Transport and the Regions, 2000b).

If the dataset is assumed to be unweighted, it is possible to inspect graphs and plots visually to examine if the reduced housing survey dataset is approximately normal. For large samples such as the 1996 EHCS (this sample has over 2,500 cases), the plots offer a balance in the numerical calculation of skew and kurtosis that exaggerates the deviation from normality. From initial histograms of the unweighted dependent variables, the log-transformed variable has a negative skew, both the logarithmic and square root transformations have some positive kurtosis, or the clustering of more cases in the centre of the distribution, and the fourth root-transformed variable has even more positive kurtosis.

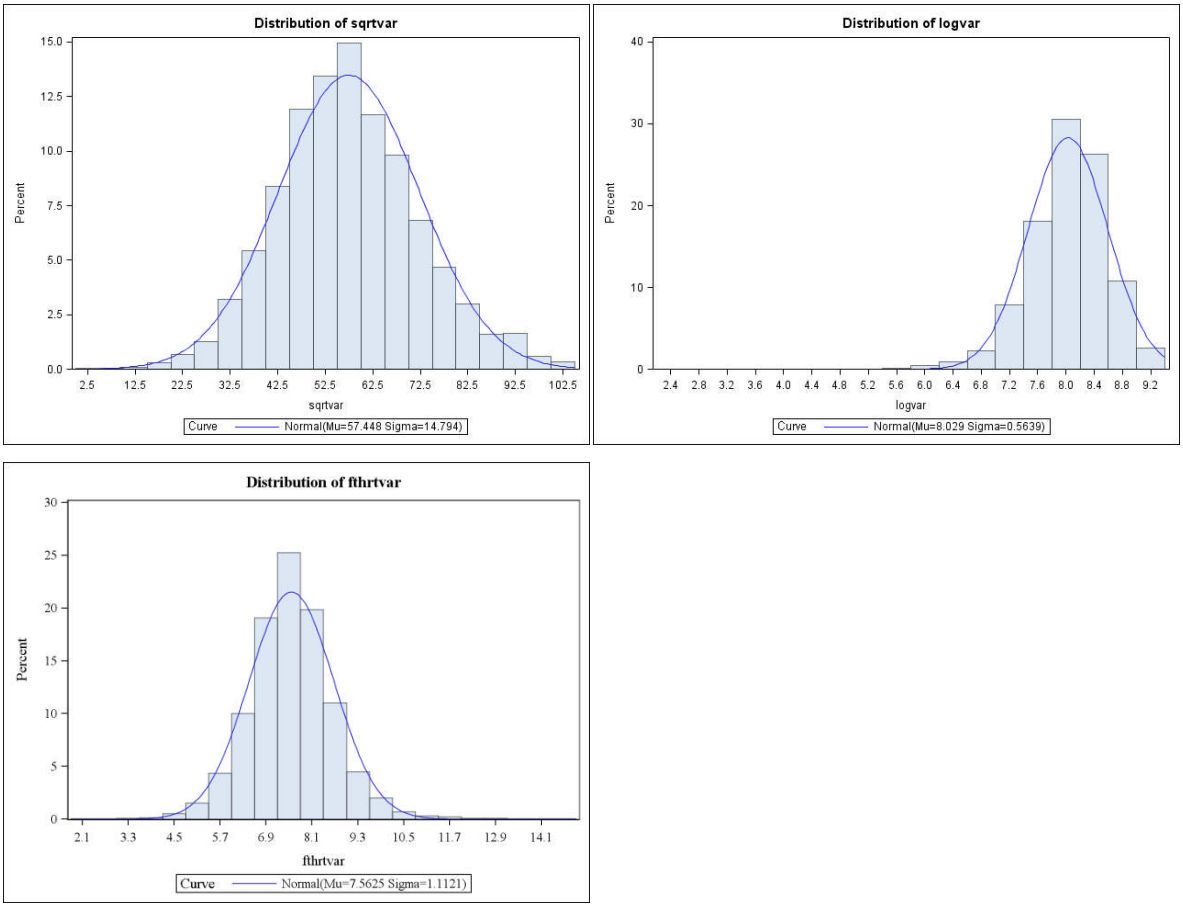


Figure 6.11: Distribution of transformations of the dependent variable: square root (top left), logarithmic (top right), fourth root (bottom left)

In order to assess the significance of the skew and kurtosis, there are visualisations available. The normal probability plot and the quantile-quantile (Q-Q) plot (Chambers, 1983) are graphical techniques for assessing whether or not a dataset is approximately normally distributed by showing the extent to which the residuals are normally distributed. The normal probability plot is particularly sensitive to deviations from normality near the centre of the distribution. A quantile-quantile plot graphs the quantiles, or points taken at regular intervals along the distribution, of a variable, against the quantiles of a normal distribution, making it easy to detect the undue influence of outliers.

This data is plotted against the theoretical normal distribution such that an approximately normal distribution in the population based on the sample forms an approximately straight line from the lower left hand to the upper right hand corner of the plot. A departure from this line indicates a departure from normality.

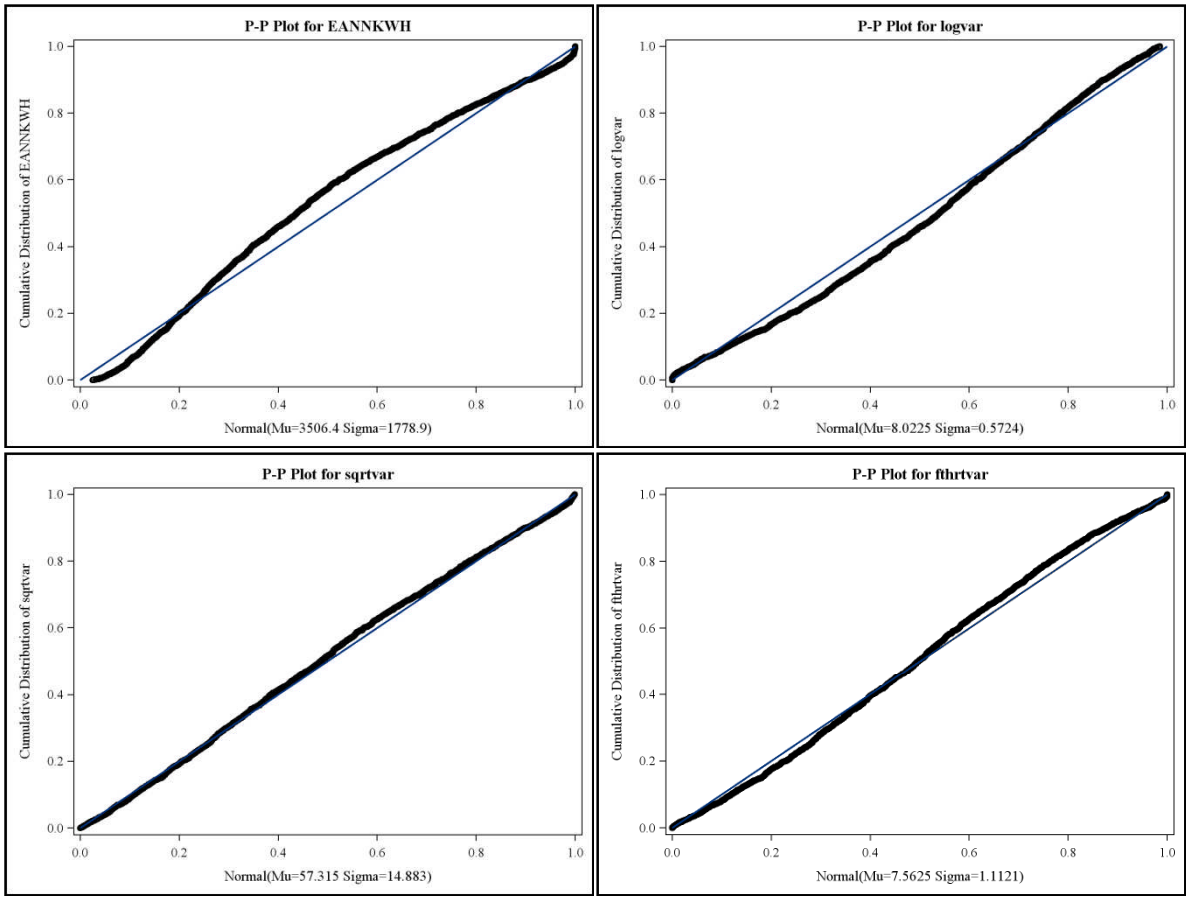


Figure 6.12: Probability plots for dependent variable, reduced dataset for untransformed (top left), natural log-transformation (top right), square root-transformed (bottom left), and fourth root-transformed (bottom right)

The normal probability plot indicates that the square root-transformed variable is approximately normal, with significant departure in the centre of the distribution in both the untransformed and the log-transformed variable, and a small departure in the fourth root-transformed variable.

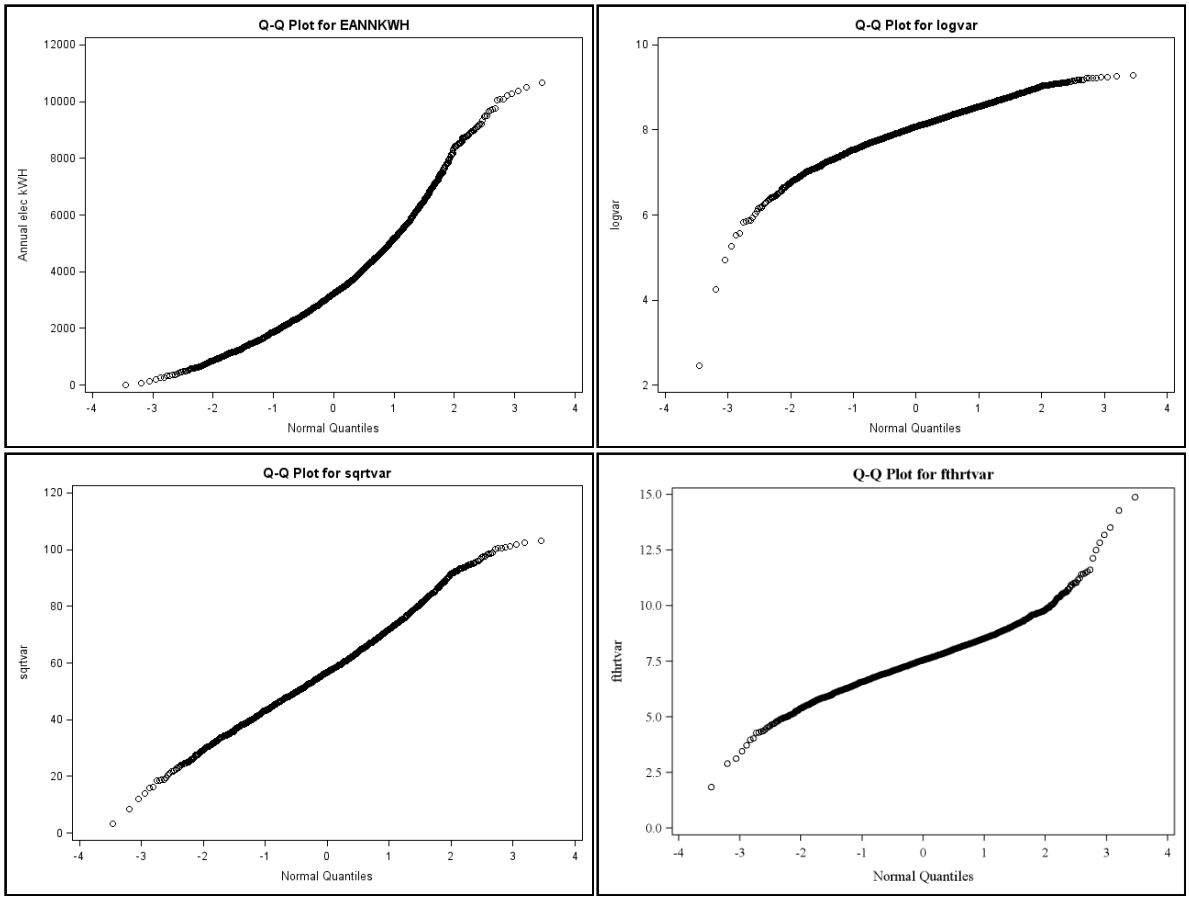


Figure 6.13: Quantile-quantile plots for dependent variable, reduced dataset for untransformed (top left), natural log-transformation (top right), square root-transformed (bottom left), and fourth root-transformed (bottom right)

The quantile-quantile plot indicates significant departures from normality for the untransformed , the fourth root-transformed, and the log-transformed variable. The plot also indicates that the square root-transformed variable is approximately normal with a slight deviation from normality indicated for very large energy users $\{\sqrt{energy} > 100\}$.

6.5.3 Conclusions from transformation and parametric tests of raw data from the 1996 English House Condition Survey

The result of the examination of the fuel sample of the 1996 English House Condition Survey is that a reduced dataset without outliers or high leverage points, transformed by a square root, has an approximately normal distribution and passes the parametric tests as an unweighted dataset based on visual inspection but backed up by the raw scores calculated using Fisher's statistical methods. Regression and analysis of variance techniques can be used on this dataset under the assumption that the sample is random, and drawn from an infinite population as opposed to a stratified sample drawn from a specific population. The reasons behind this are that the complex data collection

design is weakened sufficiently in the reduced fuel sample so that data can be examined as a random sample of the population.

6.6 Data availability for the annual option: Estimation for 2008 of participants in the 1996 English House Condition Survey

In order to prepare data for an annual option put forward for consideration in Chapter 5, the sample of homes in the 1996 English House Condition Survey (EHCS) and the population of homes in the 2001 Census should be converted to the year 2008. In this section, the 2008 Living Costs and Fuel Survey (LCFS) is used to estimate the electricity use in the year 2008 for homes surveyed in the 1996 EHCS. This will enable the direct comparison of a bottom-up housing stock model of non-heating end-use energy with the small area statistics collected on electricity consumption in all homes by the Department of Energy and Climate Change.

The 1996 EHCS has data from electricity meters that were read once a quarter. British energy customers have historically billed their customers by quarters. Before liberalisation of the electricity market in May 1999, quarterly bills in credit were the only choice outside of pre-payment slot meters given to those who had either requested them or had previously fallen behind on their bills (Which, 2011). In the 1996 EHCS, only those who were on quarterly billing cycles had meter readings on the same cycle, presumably using the arrival of an electricity bill as the prompt for the occupant to write down the current meter reading to mark the end of the measurement quarter (Department of the Environment Transport and the Regions, 2000b). The data shows some variation amongst the seasons; the researchers that created the algorithm for non-heating end-use energy in new buildings used this data to take account of seasonality.

The 2008 LCFS asks participants the amount of their last electricity bill and the value of any rebate from the previous billing cycle to produce a total electricity liability for either the last month or quarter (monthly direct debits have become more popular since the liberalisation of the energy market). These electricity bills are then converted into kilowatt-hours of energy, using data on the average regional electricity rates for credit and direct debit customers. Unfortunately, prepayment customers are again not part of the dataset as their latest electricity payment is not connected to a time period in the 2008 LCFS. All of the monthly data is extended across an entire quarter using the month before of the electricity bill as the end of the quarter (Npower, 2011, British Gas, 2011). Each quarterly bill was evenly distributed across the three months of the billing cycle.

Electricity unit costs in pounds per kilowatt-hour for the year 2008 by city and payment type were accessed via the Department for Energy and Climate Change (Department for Energy and Climate Change, 2010a). The data for each city was assigned to its government office region, with a mean taken of the unit costs if there were multiple cities in a region. All the cases in the 2008 LCFS then have their energy bill per month converted into kilowatt-hours per month using the unit cost for the home's region.

The quarterly electricity consumption data in the 1996 EHCS is also converted into monthly electricity use across the three months before the meter reading date for that quarter. Most of the participants in the fuel survey took 9 consecutive quarters of meter readings during the course of around two years. If there is repeat observation for a month in the following two years of the fuel survey, then the mean of the repeated observations is taken to represent the energy use for that month. This assumes that seasonality in electricity use for non-heating end-uses does not vary with year-to-year differences in weather for the same month.

An estimate for the electricity usage in the year 2008 can now be made for the participants in 1996 EHCS. First, this assumes that the homes in 2008 have the same composition as they had in 2008, and that the factors that explain the differences between homes are constant over time. The monthly electricity data for the 1996 homes is standardised around a mean of zero and a standard deviation of 1 for each month of the year. The 2008 estimate is then computed as:

$$E_{2008} = \sum_{m=1}^{12} E_{m,2008} \quad (73)$$

where E_{2008} is the annual electricity use for non-heating end-uses for a domestic household and $E_{m,2008}$ is the electricity use in month *m* in the year 2008. Each monthly energy use estimate $E_{m,2008}$ is computed as:

$$E_{m,2008} = \bar{E}_{m,2008} + \left(z_{m,1996}\sigma_{m,2008}\right) \quad (74)$$

where $\bar{E}_{m,2008}$ is the mean electricity use for non-heating end-uses in all domestic households surveyed in the 2008 LCFS for month *m,* $z_{m,1996}$ is the z-score for electricity use for non-heating end-uses in an individual household surveyed in the 1996 EHCS around a mean of zero and a standard deviation of 1 for month *m*, and $\sigma_{m,2008}$ is the standard deviation of electricity use for non-heating end-uses in all domestic households surveyed in the 2008 LCFS for month *m.*

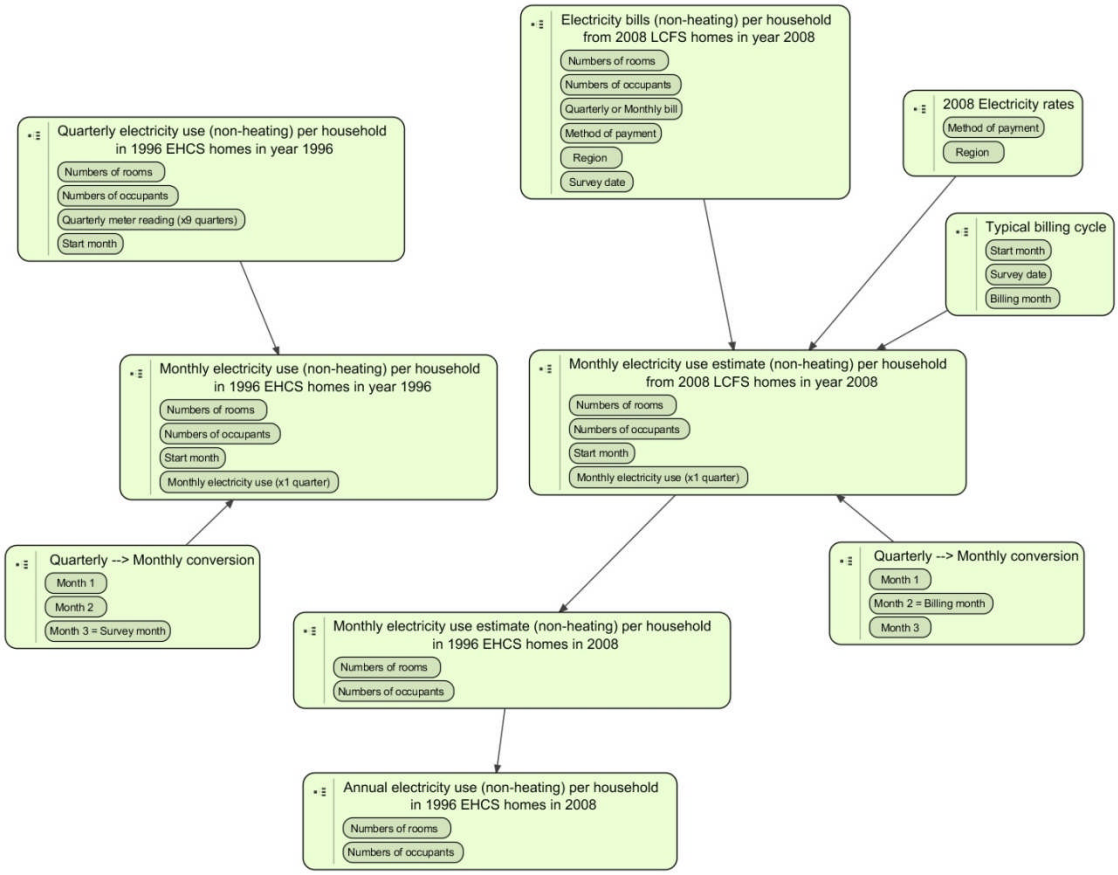


Figure 6.14: Calculation of annual electricity use in 2008 for households in the fuel sub-sample of the 1996 EHCS

The result of this process are new annual estimates of non-heating end-use energy for homes in the 1996 EHCS for the year 2008. This procedure can be repeated for subsequent expenditure surveys until the next round of housing surveys, census data, and area classifications are made available for England every ten years. For information, the monthly means and medians are shown below.

Table 6.10: Final differences between actual 1996 and estimated 2008 non-heating energy use by month (kWh/month)

Months (Jan – Dec = 1 to 12)	1996 EHCS, measured		2008, estimated	
Variable	**Mean**	**Median**	**Mean**	**Median**
ekwhmonth1	277	247	292	260
ekwhmonth2	270	241	291	264
ekwhmonth3	265	238	303	272
ekwhmonth4	267	237	309	271
ekwhmonth5	270	238	302	260
ekwhmonth6	273	241	347	288
ekwhmonth7	278	245	302	269
ekwhmonth8	285	252	319	284
ekwhmonth9	287	256	302	271
ekwhmonth10	284	253	335	292
ekwhmonth11	282	252	322	291
ekwhmonth12	282	250	343	301

6.7 Conclusions

This chapter provided evidence that the data being considered was appropriate and valid for further statistical analysis. The data ranged from individual household surveys to aggregates of geographically defined areas and area classification systems. The dependent variable of energy use in non-heating end-uses was firmly defined as electricity consumption in households not using electricity for heating. An interaction term was established as the product of two independent variables - the numbers of occupants with the number of habitable rooms.

The individual level data was examined for its ability to be analysed using statistical methods that assumed that the dependent variable passed the parametric tests. In the original fuel sample, the survey design does not permit the researcher to assume random sample selection, as a sub-sample of the survey contains grossing factors, confirmed by a comparison of the total sample with the weighted sub-sample. However, this design does not hold as firmly after the reduction in observations due to the restrictions of the dependent variable. The exclusion of outliers and high leverage points reduces the households covered using the grossing factors by around a quarter of

the total households in England after comparing distributions of the total sample with the sub-sample with weights. Therefore, research can proceed with analysis of the data either as a random sample from an infinite population or as a weighted sample representing three-quarters of the English population.

Finally, the available data was examined for demonstrating the ability for the annual option to be developed for a modelling technique. The conclusion was that there was a valid process for converting the energy use scores from 1996 to estimated values for 2008 using standardised scores from 1996 and mean values and standard deviations from 2008. However, the results are unstable as both the datasets have a positive skew, and the 2008 data only covers one electricity bill per year.

Chapter 7 - Running the models

7.1 Introduction

This section describes the running of the four model options. A prediction is made in each case for the non-heating end-use energy of the housing stock of a geographically defined area. The first two options use a simple single-level model to predict energy use with household size using the annual and decennial options outlined in Chapter 5. The second two options use a multilevel model to predict energy use by household size and area typology. Each of these models is run against the actual energy use of small statistical areas where there is a low incidence of predicted use of electricity for heating end-uses.

7.2 Single-level Model

As explained in previous chapters, the traditional, single-level model of non-heating end-use energy in dwellings has been altered to use as an interaction term of two independent variables of the numbers of rooms and the numbers of occupants of a building, both of which are directly measured in housing surveys to provide data for the model. The number of occupants in this model is, unlike BREDEM or SAP, not dependent on the physical size of the dwelling (e.g. usable floor space). This is intended to be run on the existing housing stock where both these independent variables are known.

The single-level model will produce two different algorithms, one that is based on the fuel sample of the 1996 English House Condition Survey, and a second algorithm based on a modification of the fuel sample using the 2008 Living Costs and Food Survey. These are proxies for the decennial option and the annual models presented in previous chapters.

7.2.1 First run of the linear regression model (decennial option)

The single-level model was run using the PROC REG procedure in SAS (SAS Institute, 2011). The dataset was reduced from the raw dataset by an initial removal of outliers and high leverage points as detailed in Chapter 6. The original run of the model shows the following diagnostics. The first set of diagnostics is the analysis-of-variance table:

Table 7.1: Analysis of variance, linear regression model, decennial option

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	155300	155300	1060	<.0001
Error	2288	335300	146.6		
Corrected Total	2289	490600			

The analysis of variance table contains the following (UCLA, 2011):

- Source – the variance of the dependent variable is broken down into the categories of Model, Error, and Corrected Total. The analysis of variance partitions the total variance into the variance which can be explained by the independent variables (Model) and the variance which is not explained by the independent variables (Error).
- DF - These are the degrees of freedom associated with the sources of the variance. The total variance has N-1 degrees of freedom (in this model N=2290). The model degrees of freedom correspond to the number of coefficients estimated minus 1. Including the intercept, there are 2 coefficients, so the model has 2-1=1 degrees of freedom. The Error degrees of freedom is the DF total minus the DF model, 2289 – 1 = 2288.
- Sum of Squares - These are the Sums of Squares associated with the three sources of variance, Total, Model and Error.
- Mean Square - These are the Mean Squares, the Sums of Squares divided by their respective DFs.
- F Value - This is the F-statistic: the Mean Square Model (155292) divided by the Mean Square Error (146.55761), yielding F=1059.59.
- Pr > F - This is the p-value associated with the above F-statistic. It is used in testing the null hypothesis that all of the model coefficients are 0 and have no predictive power. As the p-value is less than 0.001, then the F-statistic, and therefore the interaction term has predictive power on the electricity use of a household.

The overall model fit is assessed by the following statistics:

Table 7.2: Overall model fit, linear regression model, decennial option

Root MSE	12.11	**R-Square**	0.3165
Dependent Mean	57.11	**Adj R-Squared**	0.3162
Coeff Var	21.20		

- Root MSE - The square root of the Mean Square Error, or the standard deviation of the error term
- Dependent Mean - This is the mean of the dependent variable (the square root transformation of the annual non-heating end-use energy consumption of a household, represented by the electricity use of a household where it does not use electricity for heating end-uses).
- Coeff Var - This is the coefficient of variation, which is a unit-less measure of variation in the data. The root MSE is divided by the mean of the dependent variable, multiplied by 100: (100*(12.11/57.11) = 21.20).
- R-Square - R-Squared is the proportion of variance in the dependent variable (non-heating end-use energy) which can be explained by the independent variables (size of household measured by the number of rooms and number of occupants). This is an overall measure of the strength of association of all the independent variables with the dependent variable. As R-Squared is only 0.31, this is a first alert that there may be further outliers and high leverage points based on the residuals, as opposed to the raw score as explored in Chapter 6.
- Adj R-Sq - This is an adjustment of the R-squared that penalizes the addition of extraneous additional predictors to the model. As the sample size is large, this adjustment makes little difference to R-squared.

Thirdly, the parameter estimates in PROC REG form the algorithm that predicts the dependent variable from the independent variables:

Table 7.3: Parameter estimates, linear regression model, decennial option

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	43.98	0.4761	92.39	<.0001
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	1	1.009	0.0310	32.55	<.0001

The diagnostic statistics for this run of the linear regression model are as follows:

- Variable - This column shows the independent variables (in this case, the interaction term that represents the two independent variables of the number of rooms and number of occupants). The first refers to the model intercept or the height of the regression line when it crosses the Y axis. In other words, this is the predicted value of non-heating end-use energy when all other variables are 0.
- Label - This column gives the label for the variable.
- DF - This column give the degrees of freedom associated with each independent variable. All continuous variables, such as the interaction term, have one degree of freedom.
- Parameter Estimates - These are the values for the regression equation for predicting the dependent variable from the independent variable. For the interaction term, this is its regression coefficient with a value of 1.01.
- Standard Error - The standard error of the coefficient.
- t Value - These are the t-statistics used in testing whether the coefficient generated as a parameter estimate is significantly different from zero.
- Pr > |t|- The p-value used in testing the null hypothesis that the coefficient (parameter) is 0.

Therefore, the initial regression equation for predicting non-heating end-use energy from the size of the household is (to three significant figures):

$$\sqrt{e} = (1.01mn + 44.0) \quad (75)$$

where *e* is annual household non-heating end-use energy in kilowatt-hours of electricity, *m* is the number of rooms in the household, and *n* is the number of occupants of the household. This formula is squared on both sides to give:

$$e = (1.01mn + 44.0)^2 = 1.02m^2n^2 + 88.8mn + 1940 \quad (76)$$

The next series of diagnostics of the fit are as follows:

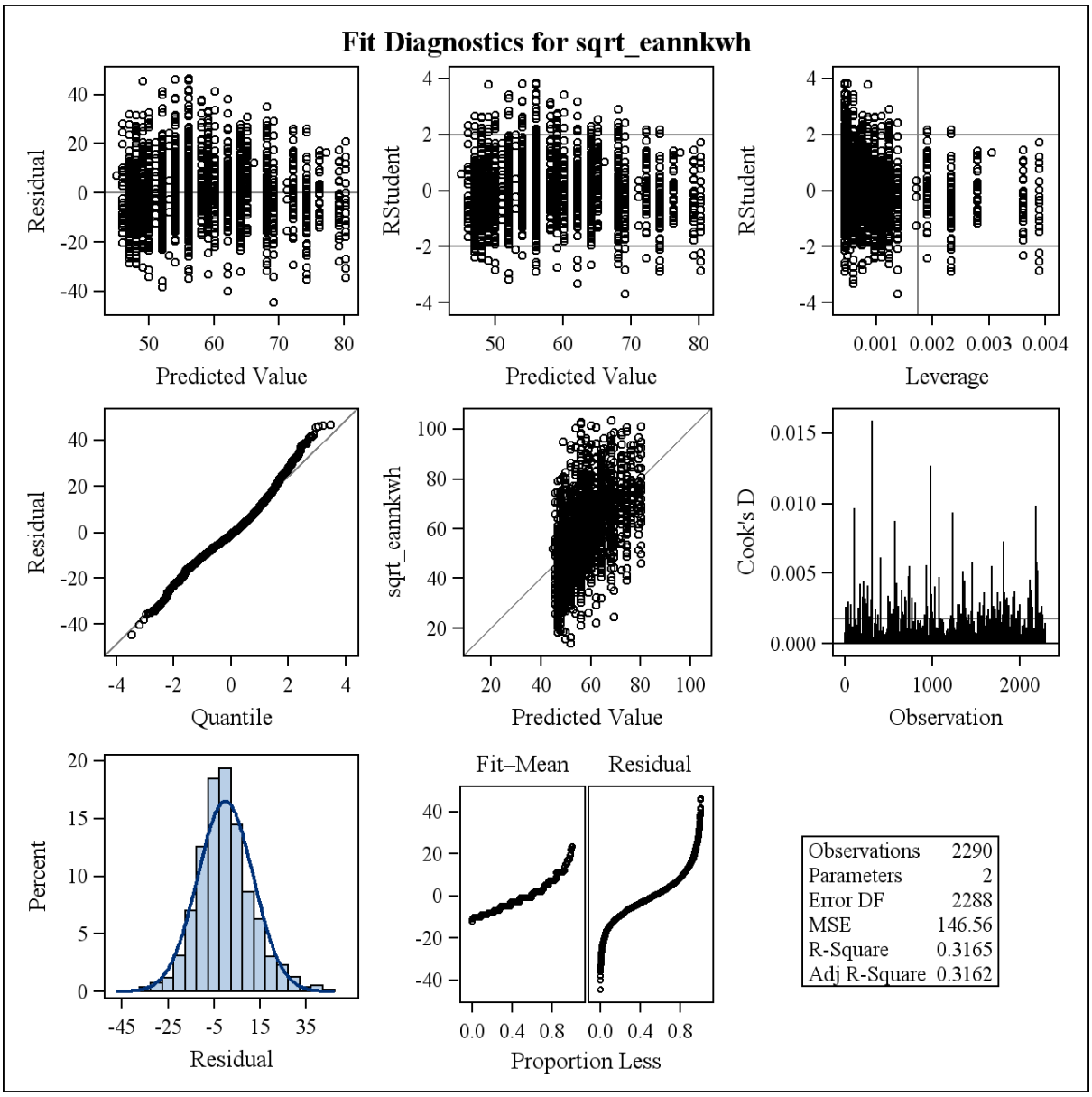


Figure 7.1: Fit diagnostics for the interaction term, linear regression model, decennial option

There are issues that show up in three key graphs: the studentised residuals (Rstudent) against both the predicted values (dependent variable) and the leverage (interaction term). A studentised residual is the value of the residual divided by its estimated standard deviation. Values above 2 or below 2 are therefore estimated to be outliers or high leverage points in a large sample set. A histogram of the residuals shows that the residuals are approximately normal and centred around zero. Another assumption is that the residuals are heteroskedastic, or there is no discernable pattern of residuals. There is a mild pattern, as residuals grow for predicted values of 50-60. The result of

this analysis is that there is a group of outliers around a predicted value of 56 that should be examined further and considered for removal.

7.2.2 Second run of the linear regression model (decennial option)

A second run of the linear regression model can be made with a slightly reduced set of observations with the intention of increasing the reliability of the linear model. If all studentised residuals greater than 2 and less than 2 are removed, the number of observations goes down by 6% to 2156. Almost all of these represent values greater than 2, where the algorithm predicts less than the measured value. The following regression diagnostics are found:

Table 7.4: Analysis of variance, linear regression model, decennial option, second run

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	157700	1577	1654	<.0001
Error	2154	205400	95.36		
Corrected Total	2155	363200			

Table 7.5: Overall model fit, linear regression model, decennial option, second run

Root MSE	9.766	**R-Square**	0.4343
Dependent Mean	56.34	**Adj R-Sq**	0.4341
Coeff Var	17.33		

Table 7.6: Parameter estimates, linear regression model, decennial option, second run

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	42.72	0.3954	108.0	<.0001
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	1	1.053	0.0259	40.67	<.0001

The R-squared value, or the proportion of variance explained in the dependent variable by the interaction term, has gone up substantially from 0.32 to 0.43. The removal of these outliers with large residuals also has resulted in a more random and normal distribution of residuals:

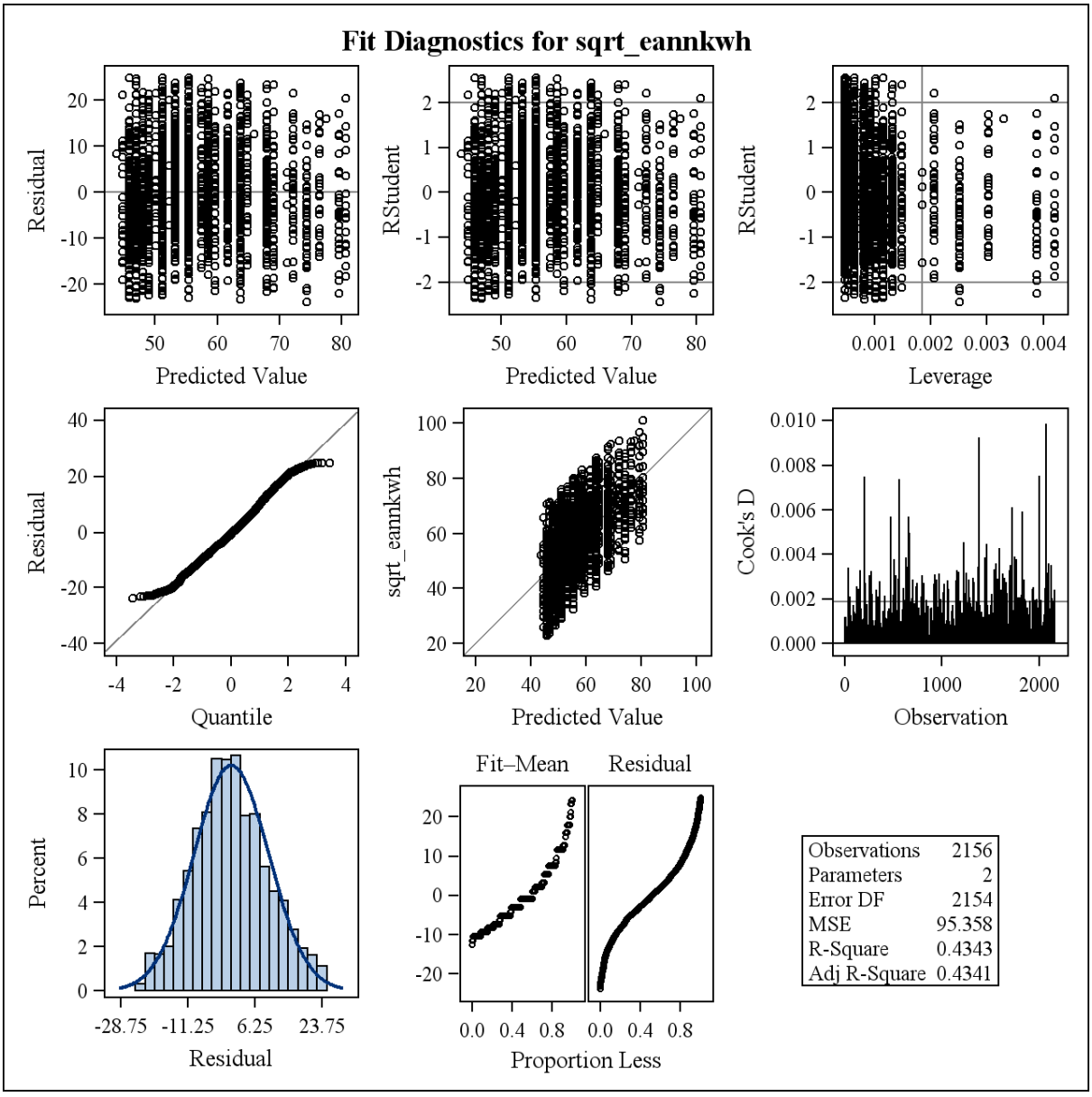


Figure 7.2: Fit diagnostics for the interaction term, linear regression model, decennial option, second run

The resulting algorithm for the second run of the linear regression is:

$$e = (1.05mn + 42.7)^2 = 1.10m^2n^2 + 89.7mn + 1820 \quad (77)$$

7.2.3 Running the linear model with the estimation of 2008 energy use of homes in the 1996 English House Condition Survey (annual option)

Previously, a method was developed for estimating the 2008 energy use of homes that were included in the 1996 English House Condition Survey. The same regression analysis was performed on the data, with the following regression diagnostics for the first run:

Table 7.7: Analysis of variance, linear regression model, annual option

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	125700	125700	978.7	<.0001
Error	2119	272200	128.44		
Corrected Total	2120	397900			

Table 7.8: Overall model fit, linear regression model, annual option

Root MSE	11.33	**R-Square**	0.3159
Dependent Mean	56.81	**Adj R-Sq**	0.3156
Coeff Var	19.95		

Table 7.9: Parameter estimates, linear regression model, annual option

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	44.02	0.4771	92.26	<.0001
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	1	1.012	0.03234	31.28	<.0001

And the second run:

Table 7.10: Analysis of variance, linear regression model, annual option, second run

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	130900	130900	1504	<.0001
Error	2000	174100	87.04		
Corrected Total	2001	305000			

Table 7.11: Overall model fit, linear regression model, annual option, second run

Root MSE	9.329	**R-Square**	0.4292
Dependent Mean	56.11	**Adj R-Sq**	0.4289
Coeff Var	16.63		

Table 7.12: Parameter estimates, linear regression model, annual option, second run

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	42.68	0.4043	105.6	<.0001
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	1	1.063	0.02742	38.78	<.0001

The parameter estimates are not very different from the original energy consumption data from the 1996 English House Condition Survey. The coefficients, or the effect of household size are slightly greater in 2008 than in 1996, with the intercept slightly lower. The R-squares are nearly identical in the 2008 estimates of energy consumption compared with the measured consumption in 1996 of the same houses.

The value for the 2008 estimate of energy use per household is:

$$\sqrt{e} = (1.06mn + 42.7) \quad (78)$$

This formula is squared on both sides to back-transform from the square root to:

$$e = (1.06mn + 42.7)^2 = 1.12m^2n^2 + 90.5mn + 1820 \quad (79)$$

7.2.4 Observations and preliminary conclusions

The resulting algorithm was not entirely expected, as the parameter estimate produced a positive coefficient, where previous studies predicted a negative term to diminish the effect of growing household size on non-heating end-use energy. Instead, the square root transformation of the dependent variable has produced an algorithm of the structure $(a + b)^2$ where the effect of household size increases as the numbers of rooms and occupants grows to a large number. Previous attempts to estimate non-heating end-uses in the English context have concluded that there should be a maximum value in the model (Anderson et al., 1996, Anderson et al., 1985a, Henderson and Shorrock, 1986b). The values associated with large households should be looked at closely in the evaluation of any model of energy use.

7.3 Validation of the single-level model against actual 2008 energy use levels in small areas

Each lower layer super output area has two datasets that relate to the size of the household. The first of these is a dataset that details the number of households in a census area by the number of habitable rooms that they contain (Office of National Statistics, 2005). The second is the number of people living in a household (Office for National Statistics, 2005b). A cross-tabulation of the two is available with the following categories for household size – 1 person, 2 persons, 3 to 4 persons, and 5 or more persons, and 1 room, 2 rooms, 3 or 4 rooms, 5 or 6 rooms, and 7 or more rooms.

Analysis of the 1996 English House Condition Survey reveals that the following weighted means are found for these categories. For the largest category, medians were considered instead of means because of the skew caused by larger homes. In this case, the value of the median is the lower bound because of the preponderance of these homes in the population. The final representative average was placed between the mean and the median to ensure that the size of the household was not overestimated in a lower output area.

Table 7.13: Selection of a representative average from the fuel sub-sample of the 1996 EHCS

Category	Weighted Mean (1996 EHCS)	Weighted Median (1996 EHCS)	Representative Average (1996 EHCS)
3 to 4 rooms	3.73	N/A	3.73
5 to 6 rooms	5.52	N/A	5.26
7 or more rooms	7.61	7.00	7.30
3 to 4 people	3.50	N/A	3.50
5 or more people	5.52	5.00	5.26

An estimate for non-heating end-use energy in 2001 can be made for each Lower Layer Super Output Area from the resulting cross-tabulation for each area:

Table 7.14: Cross-tabulation proforma to estimate each LLSOA's non-heating end-use energy consumption in 2001 using Census data

	1 room	2 rooms	3.73 rooms	5.52 rooms	7.30 rooms
1 occupant					
2 occupants					
3.50 occupants					
5.26 occupants					

The cross-tabulation was updated for 2008 using the estimate for households made by the DECC and the BRE for the total number of households in 2008 as part of fuel poverty research using data from the 2007 and 2008 English Housing Survey (Department of Energy and Climate Change and BRE, 2010). The new number of households was assumed to have the same distribution of household sizes represented by the number of rooms and the number of occupants as in 2001. This enables the estimate of energy use for the year 2008 for each household to match the same year as the measured electricity (ordinary charge) energy use in a Lower Layer Super Output Area.

Using these cross-tabulations for each census area, the total electricity use for these areas is estimated to be less than 2% from the modelled electricity use of all census areas. An examination of the differences by area classification supergroup name shows that there were some differences between the predicted and actual electricity use for non-heating end-uses, but the maximum error is still just over 5%.

Table 7.15: Difference between estimated and actual non-heating energy use in 2008 by LLSOA ONS area classification supergroup

ONS 2001 Area Classification Supergroup Name	Number of LSOAs with central heating >95%	Model estimate of electricity use in 2008 (kWh)	Actual electricity use recorded in 2008 (kWh)	Difference between estimate and actual use (kWh)	Difference between estimate and actual use (percent)
Countryside	673	1,552,265,666	1,604,585,375	-52,319,710	-3.3%
Disadvantaged Urban Communities	1,933	3,912,787,843	3,736,427,124	38,540,686	1.0%
Miscellaneous built up areas	3,068	6,690,162,789	6,356,415,994	329,603,350	5.2%
Multicultural City Life	1,479	2,942,719,802	2,836,110,254	105,202,641	3.7%
Professional City Life	827	1,837,506,949	1,909,557,762	-75,538,266	-4.0%
Urban Fringe	1,149	2,639,688,748	2,631,764,847	-56,820,327	-2.2%
White Collar Urban	1,221	2,653,822,364	2,595,908,579	34,494,415	1.3%
TOTAL	**10,350**	**22,228,954,160**	**21,670,769,936**	**323,162,789**	**1.5%**

7.4 Multilevel model

The 1996 English House Condition Survey was matched with the area classification data in a multilevel model as described in previous sections. As discussed in earlier chapters, this multilevel model uses the both the interaction term of the household size of occupants and the number of rooms of individual households measured in the 1996 English House Condition Survey (x_{ij}) and the interaction mean household size for each area classification at both the supergroup and group levels as measured in the 2001 United Kingdom Census ($\bar{x}_j$) as the two terms that are composed of independent variables.

The classifications of valid cases at the supergroup level in the 1996 English House Condition Survey dataset are as follows:

Table 7.16: Valid cases in the fuel sub-sample of the 1996 EHCS by LLSOA ONS area classification supergroup

SUPERGROUP_NAME	Number
Countryside	72
Disadvantaged Urban Communities	361
Miscellaneous built up areas	358
Multicultural City Life	210
Professional City Life	143
Urban Fringe	169
White Collar Urban	296

The classifications of valid cases at the group level in the 1996 English House Condition Survey dataset are as follows:

Table 7.17: Valid cases in the fuel sub-sample of the 1996 EHCS by LLSOA ONS area classification group

SUPERGROUP_NAME	GROUP_NAME	Number
Countryside	Countryside Communities	6
Countryside	Farming and Forestry	9
Countryside	Rural Economies	57
Disadvantaged Urban Communities	Blue Collar Urban Families	183
Disadvantaged Urban Communities	Struggling Urban Families	178
Miscellaneous built up areas	Resorts and Retirement	75
Miscellaneous built up areas	Small Town Communities	122
Miscellaneous built up areas	Suburbia	70
Miscellaneous built up areas	Urban Terracing	91
Multicultural City Life	Multicultural Inner City	81
Multicultural City Life	Multicultural Suburbia	79
Multicultural City Life	Multicultural Urban	50

SUPERGROUP_NAME	GROUP_NAME	Number
Professional City Life	Educational Centres	23
Professional City Life	Mature City Professionals	77
Professional City Life	Young City Professionals	43
Urban Fringe	Affluent Urban Commuter	79
Urban Fringe	Urban Commuter	90
White Collar Urban	Mature Urban Households	116
White Collar Urban	Well off Mature Households	124
White Collar Urban	Young Urban Families	56

An analysis of variance was performed to investigate the effect of area classifications on non-heating end-use energy. This was done using PROC MIXED in SAS (SAS Institute, 2011) using maximum likelihood estimation (Hartley and Rao, 1967) to compare the models in an analysis of variance. In both the supergroups (DF=6, F=8.25, p<.001) and groups (DF=19, F=3.42, p<.001) there was a significant effect of group membership on the dependent variable (square root transformed non-heating end-use energy) within the fuel sample of the 1996 English House Condition Survey. Therefore the multilevel model is a good way of explaining more of the variance of non-heating end-use energy than simply the size of individual households.

As explained in previous chapters, the overall fit of a multilevel model is tested using the chi-square likelihood ration test, which reports the -2 log likelihood, or the likelihood ratio statistic. The likelihood-ratio statistic is based on the comparison between the frequencies of the dependent variable as observed with those the predicted frequencies of the dependent variable in the model. This statistic is:

$$L\chi^2 = 2\sum observed_{ij} \ln\left(\frac{observed_{ij}}{model_{ij}}\right) \quad (80)$$

where *i* and *j* are the rows and columns of a contingency table of the number of coefficients that refer to independent variables (the coefficients to the interaction term for individual household size in the survey and the interaction term of the mean household size for an area classification) *i* and covariance parameters (classification types and their residual) *j*. If the model changes the number of these parameters and coefficients, a new likelihood statistic should be computed. If the number

goes down, the model may have more predictive power. The difference between the two models should be tested against the critical values of the chi-square distribution against the difference in the number of degrees of freedom which are the total of parameters and coefficients above. For example, 1 degree of freedom difference would require a difference in the likelihood statistic over 6.63 if *p*=.01 or 3.84 for *p*=.05 to reject the null hypothesis.

Log-likelihood will always go down if there are more predictors, so if there is more than one degree of freedom difference between two models. An alternative option is to rely on the Akaike information criterion (AIC) which corrects for the additional predictors in the model. AIC is calculated as the likelihood-ratio statistic plus twice the number of predictors. As in the likelihood statistic, a lower number equals a better fit, though because of the correction factor, there is no requirement for the difference between models to meet a critical value of the chi-square distribution (Field and Miles, 2010b).

The test of the effectiveness of the multilevel model is run through the following progression using the PROC MIXED procedure in SAS:

1. Analysis of covariance of the interaction term at the individual level (x_{ij}) and the mean interaction term at the group level ($\bar{x}_j$) without taking into account the data structure
2. Factoring in the data structure by allowing random intercepts
3. Factoring in the data structure by allowing random intercepts and slopes

As stated earlier, both the decennial and annual options are explored. The annual option is based on the estimate of the homes in the 1996 English House Condition Survey in 2008 using data from the 2008 Living Costs and Food Survey. The mean interaction terms of census areas in the 2001 Census are assumed to be the same in the same census area in 2008.

7.4.1 Running of the multilevel model - unconditional means

First, one needs to examine the question of how different area classification supergroups vary in their mean household size as measured by the interaction term of the number of rooms with the numbers of occupants. This model is called the unconditional means model (Singer and Willett, 2003) and is represented by:

$$\sqrt{e_{ij}} = \beta_{0j} + \varepsilon_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij} \quad (81)$$

where $\sqrt{e_{ij}}$ is the square-root transformed dependent variable of non-heating end-use energy, β_{0j} is the intercept for the area classification supergroup *j*, γ_{00} is the overall intercept, or mean value of household size, u_{0j} is the random deviation of supergroup *j* from the overall mean, and ε_{ij} is the

random error associated with household *i* in supergroup *j.* The model assumes the residuals σ^2 and τ_{00} of ε_{ij} and u_{0j} are both normally distributed around a mean of zero.

Both the interaction term for household size and the mean interaction term for each supergroup have been centred around the grand mean for England. The 2001 Census states that the average number of occupants in a household was 2.36 and the average number of rooms in a household was 5.33.

The PROC MIXED procedure in SAS (SAS Institute, 2011) produces covariance parameter estimates for the random effects of the model, resulting in an estimated value of $\sigma^2 = 168$ and $\tau_{00} = 5.28$.

Table 7.18: Covariance parameter estimates of random effects by supergroup, unconditional means of the interaction term

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	SUPERGROUP_NAME	5.283	3.518	1.500	0.0666
Residual		168.5	5.952	28.30	<.0001

Using $p = 0.1$ as the standard for statistical significance, the hypothesis test suggests that both of these variance components are significantly different from zero at both the group (*p*=.067) and at the individual (*p*<.001) level. However, the Wald statistic *Z* in this procedure has been discussed in previous papers to be unreliable for random effects (Field and Miles, 2010b, Singer, 1998). The above covariance parameter estimates imply that supergroups do differ in their average annual household non-heating end-use energy, but unsurprisingly, it also suggests that there is even more variation to be found between households within each of these supergroups.

The overall mean γ_{00} is a fixed effect of the average group-level annual household energy use in this sample of census areas and is estimated to be at 57.0 (or 3250 kwh per annum). The hypothesis test suggests that this effect is significantly different from zero.

Table 7.19: Solution for fixed effects, unconditional means of the interaction term

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	57.05	0.9415	6	60.59	<.0001

The intraclass correlation, or the measurement of how strongly units in the same group resemble one another, is defined as

$$\hat{p} = \frac{\tau_{00}}{\sigma^2+\tau_{00}} = \frac{5.28}{5.28+168} = .030 \quad (82)$$

Therefore households inside the same supergroup are 3% more similar in terms of their non-heating end-use energy than they are to households outside of their supergroup. This model has a likelihood ratio statistic of 12830, which serves as a baseline value for the usefulness of the model.

7.4.2 Unconditional means model for ONS area classification groups and government office regions

The two other viable group options, the ONS area classification groups (GROUP_NAME), and the government office regions (GOR_CODE), were also investigated:

Table 7.20: Covariance parameter estimates of random effects by group and government office region, unconditional means of the interaction term

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	GROUP_NAME	4.543	2.107	2.160	0.0155
Residual		168.0	5.953	28.22	<.0001
Intercept	gor_code	2.249	1.442	1.560	0.0595
Residual		166.3	5.076	32.77	<.0001

This output suggest that the variance at the group level is significantly different from zero in ONS area classification groups and between regions. Although the Wald statistic is unreliable, the higher value of the statistic for ONS groups suggests that there could be more explainable variation remaining after taking account of group membership than between ONS supergroups or regions. As expected, the intraclass correlations are slightly smaller for ONS groups (.026) and much smaller for regions (.013). As the regional Wald statistic suggests that a similar amount of explainable variation remained as for supergroups, regions were no longer considered as a viable option. However, the groups are considered a viable option to explore. The ONS groups, however, are viable and have the advantage of having more groups (20) and therefore should be examined alongside the supergroups.

7.4.3 Including the effects of group-level predictors

The "empty" or unconditional means model is a starting point to which more complex models can be compared. The next step is to include group-level predictors. The *centred* mean household size is added by supergroups that meet the same criterion as the validation set (>95% of households with central heating).

The model of the transformed dependent variable as a function of group-level annual household non-heating end-use energy, again, is:

$$\sqrt{e_{ij}} = \beta_{0j} + \varepsilon_{ij} \quad (83)$$

But instead of the intercept of a group β_{0j} being equal to the overall mean and a group-level deviation, this step includes the mean household size of each supergroup $\bar{x}_j$ as well as group-level deviations u_{0j}. The mean household size of the supergroup is centred around the grand mean value for the interaction term of 12.58. Therefore the group intercept formula is:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\bar{x}_j + u_{0j} \quad (84)$$

The model for the transformed dependent variable is now expressed as:

$$\sqrt{e_{ij}} = \left[\gamma_{00} + \gamma_{01}\bar{x}_j\right] + \left[u_{0j} + \varepsilon_{ij}\right] \quad (85)$$

The first bracket represents fixed effects which are applied to all individual households *i* – a rough equivalent to regression coefficients discussed in the single level model. The second bracket represents random effects: first, the variation in the intercepts between supergroups and second, the variation within these groups.

Using PROC MIXED to run a multilevel model using group-level predictors, the following table is the output of the estimate of fixed effects:

Table 7.21: Solution for fixed effects including supergroup-level predictors

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	56.96	0.7362	5	77.37	<.0001
super_mean	0.9072	0.4142	5	2.190	0.0801

$\gamma_{00} = 57.0$ is highly significantly different from zero (p<.001). This represents the transformed non-heating end-use energy in a supergroup (3250 kWh) that has the same average household size interaction term as the average for all of England.

The second fixed effect $\gamma_{01} = 0.907$ explains the relationship between energy use for non-heating end-uses and the average household size in the supergroup. This means that a 1-unit increase in the mean interaction term increases the energy use of its households by .9 units of the transformed dependent variable. The *t-statistic* of the standard error is 2.19 (p=0.080), which indicates that the null hypothesis is just about rejected that there is no relationship between the average household size of a supergroup and the energy use of its households.

An approach of measuring how much of the variation in supergroup mean energy use is explained by the mean household size is to compute how much the group-level variance component τ_{00} has diminished from the "empty" unconditional means model to the model that includes mean household size as a group-level predictor (Raudenbush and Bryk, 2002):

$$v = \frac{\tau_{00,initial} - \tau_{00,second}}{\tau_{00,initial}} = \frac{5.28 - 2.90}{5.28} = .451 \quad (86)$$

This statistic should be interpreted as saying that 45.1% of the explainable variation in supergroup mean energy use is explained by mean household size of the supergroups. There likely is not more of the variation mean energy use of a supergroup that remains to be explained as the Wald statistic of the group level variance component τ_{00} is 1.24 (p=.108). It is likely that the null hypothesis that τ_{00} is 0 cannot be rejected.

Table 7.22: Covariance parameter estimates of random effects including supergroup-level predictors

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	SUPERGROUP_NAME	2.898	2.345	1.240	0.1082
Residual		168.4	5.953	28.30	<.0001

7.4.4 Including the effects of group-level predictors of ONS area classification groups

This section replicates the above analysis, but uses ONS area classification groups instead of supergroups. The null hypothesis that there is no difference between groups is again rejected. However, unlike the supergroups, the null hypothesis that the group level variance component τ_{00} is 0 was rejected (z=1.94, p=.0261). This suggests that there is additional variation present between groups. Therefore, the ONS area classification groups and not the supergroups become the primary focus of investigation of the relationship between group membership, household size at the individual and group level and use of non-heating end-use energy at the individual and group level.

Table 7.23: Solution for fixed effects including group-level predictors

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	56.79	0.5703	18	99.58	<.0001
group_mean	0.4821	0.2759	18	1.750	0.0977

Table 7.24: Covariance parameter estimates of random effects including group-level predictors

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	GROUP_NAME	3.751	1.932	1.940	0.0261
Residual		168.1	5.956	28.22	<.0001

7.4.5 Including only the effects of individual household size into the model

Another option is to insert an individual-level predictor variable – individual household size – into the unconditional means model. This results in a model as

$$\sqrt{e_{ij}} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij} \quad (87)$$

which introduces an additional fixed effect or intercept per group γ_{10} and a random effect or slope per group u_{1j} to the model associated with household size as measured in $\underline{i}$ households demarcated in groups j x_{ij}. This means that non-heating end-use energy is related not only to the individual household size, but that this relationship between energy use and household size varies across different types of neighbourhoods in cities, towns, and country. Now that the intercepts and slopes are allowed to vary across groups, there is no longer just variance of the intercept and slope τ_{00} and τ_{01},, but covariance components that represent the correlation between intercepts τ_{10}, and between slopes τ_{11}.

However, this model needs to be altered to allow a correct interpretation of the coefficient β_{0j}. Across the housing sample, the centred score for household size has a mean of zero as it is the grand mean centred around the average household size for England. Therefore, β_{0j} represents the average energy use for a household of average size across the entire sample of the English House Condition Survey. It does not correspond to the average energy use of households in group *j* when controlling for individual household size. If more predictors are added to a multilevel model, the researcher would aim to see how the conditional group means relate to other predictor variables. In order to interpret β_{0j} more meaningfully and to allow a model with both individual-level and group-level independent variables, centred individual household sizes should be transformed again by

centring around the group mean. Therefore, a multilevel model using only an individual-level predictor is as follows:

$$\sqrt{e_{ij}} = \beta_{0j} + \beta_{1j}(x_{ij} - \overline{x_j}) + \varepsilon_{ij} \quad (88)$$

This can be separated into fixed effects γ and random effects *u* that can be estimated as follows:

$$\sqrt{e_{ij}} = \gamma_{00} + u_{0j} + (\gamma_{10} + u_{1j})(x_{ij} - \overline{x_j}) + \varepsilon_{ij} =$$

$$\left[\gamma_{00} + \gamma_{10}(x_{ij} - \overline{x_j})\right] + \left[u_{0j} + u_{1j}(x_{ij} - \overline{x_j}) + \varepsilon_{ij}\right] \quad (89)$$

Using PROC MIXED, the fixed effects on the average energy use of 2001 ONS Area Classification groups controlling for individual household size was calculated as follows:

Table 7.25: Solution for fixed effects controlling for the number of occupants

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	55.76	0.6364	19	87.62	<.0001
hsize	1.046	0.03027	1588	34.55	<.0001

These results indicate that the estimated average group mean annual household non-heating energy use γ_{00} is 55.76 and the average slope representing the relationship between individual household size and annual non-heating energy use γ_{10} is 1.046. The standard errors are very small, which results in large *t-statistics* and low *p-values.* Therefore, there is a significant relationship between individual household size and non-heating end-use energy. This is not surprising because the simple single-level model built from ordinary least squares linear regression estimation techniques came to a similar conclusion.

The following covariance estimates tell the researcher how much the fixed effects vary across groups. These are:

Table 7.26: Covariance parameter estimates of fixed and random effects controlling for the number of occupants including group-level predictors

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	GROUP_CODE	6.319	2.623	2.410	0.0080
UN(2,1)	GROUP_CODE	-0.1055	0.09154	-1.150	0.2491
UN(2,2)	GROUP_CODE	0.000754	0.006424	0.1200	0.4533
Residual		94.37	3.377	27.94	<.0001

These statistics are estimates of variances. 6.319 is the estimate of the variability of intercepts, 0.000754 is the estimate of the variability of slopes, -0.1055 is the estimate of the covariance between the intercepts and the slopes, and 94.37 is the estimate of the variance between households. Standard errors and tests of the null hypothesis tell the researcher:

- The null hypothesis that the intercepts did not differ between groups, or there is no difference between the mean energy use of groups was rejected (effect size 6.31, p<0.01). This means the average energy use of groups is different after controlling for individual energy use levels.
- However, there is little correlation between intercepts and slopes (covariance component estimate -0.1055, p=0.249). This means there is no evidence that the effects of individual household size on non-heating end-use energy differ with average non-heating energy use of households in each ONS area classification group.
- The null hypothesis that there is no variance between slopes was not rejected (slope variance component estimate 0.00075, p=0.453). This means that there is no evidence that there are differences in the rate of growth of energy use levels as individual energy use.
- There is high variability of household energy use levels within groups (variance component 94.36, p<.0001). This means that differences in household sizes explain the within-group variation in energy use. This should be compared with the unconditional model estimate of 168.03. The result of the addition of variable individual household sizes therefore explains $\frac{168.03-94.36}{168.03}$=44% of the explainable variation within groups.

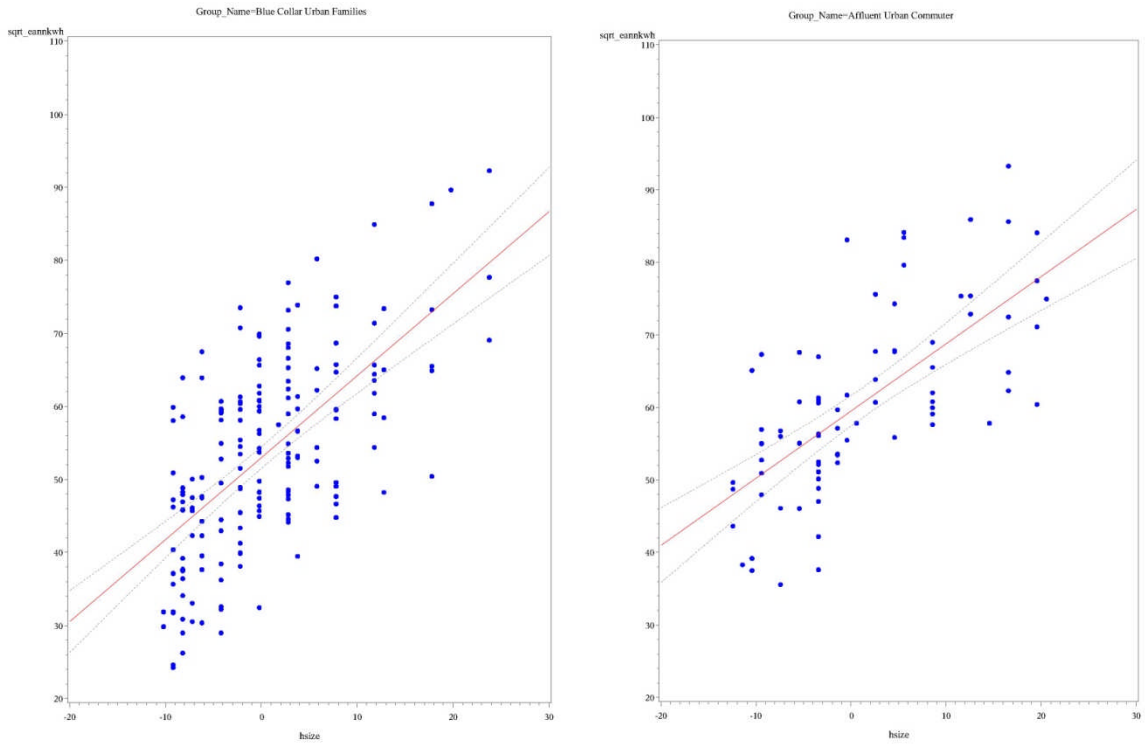


Figure 7.3: The multilevel tests rejected the null hypotheses that there was no difference between the intercepts of a linear regression of energy use against household size, but could not reject the null hypothesis that there is no difference between the slopes. This is illustrated using two example ONS groups.

7.4.6 Testing at the supergroup level

At the supergroup level , the model does not converge to a final estimate of annual household non-heating end-use energy as the number of maximum likelihood evaluations exceeded 30.

7.4.7 Including both individual-level and group-level predictors in the multilevel model

After specifying and interpreting separate models with just an individual-level and a group-level predictor, a model can be built and interpreted containing variables at both levels. First, a model containing only individual household size and average household size of ONS area classification groups is considered. The two predictor variables can be added to the model as follows:

$$\sqrt{e_{ij}} = \beta_{0j} + \beta_{1j}(x_{ij} - \bar{x_j}) + \varepsilon_{ij} \quad (90)$$

Where β_{0j} and β_{1j} contain group-level predictors and intercepts as in the previous section where group-level predictors were included in the model. Therefore:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\bar{x}_j + u_{0j}, \text{ and} \quad (91)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\bar{x}_j + u_{1j}. \quad (92)$$

This results in the following formula with fixed effects γ and random effects u as parameters that can be estimated using multilevel modelling and the PROC MIXED procedure:

$$\sqrt{e_{ij}} = \gamma_{00} + \gamma_{01}\bar{x}_j + u_{0j} + (\gamma_{10} + \gamma_{11}\bar{x}_j + u_{1j})(x_{ij} - \bar{x}_j) + \varepsilon_{ij} = \gamma_{00} + \gamma_{10}(x_{ij} - \bar{x}_j) + \gamma_{11}\bar{x}_j(x_{ij} - \bar{x}_j) + u_{0j} + u_{1j}(x_{ij} - \bar{x}_j) + \varepsilon_{ij} \quad (93)$$

where γ_{00} is the overall mean energy use of all households, γ_{01} is the fixed effect of mean household size of groups on energy use, γ_{10} is the fixed effect of individual household size on energy use, and γ_{11} is the fixed effect of the interaction of group mean household size controlling for individual household size on energy use, u_{0j} is the random variance between group mean energy use levels, u_{1j} is the random variance between group mean energy use levels controlling for individual energy use, and ε_{ij} is the random variance of individual energy use levels, $\bar{x}_j$ is the grand mean centred mean household sizes of group *j*, and x_{ij} is the individual household size of household *i.*

The same types of assumptions need to be held as before where the random variances are centred at zero with an approximately normal distribution.

The results of the model for fixed effects are as follows:

Table 7.27: Solution for fixed effects controlling for the number of occupants at the individual and group level

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	55.86	0.3918	18	142.6	<.0001
group_mean	1.103	0.1916	18	5.760	<.0001
hsize	1.052	0.03005	1587	35.01	<.0001
group_mean*hsize	-0.02121	0.01499	1587	-1.410	0.1574

The fixed effects above mean that individual-level household size level and group-level mean household size are both associated with non-heating end-use energy in households. However, there is no interaction between group mean household size and individual household size. This means that the linear relationship built between household size as the predictor and non-heating end-use

energy use as the outcome is somewhat, but not very different (built out of two components of household size) from the simple single-level model.

The covariance parameters below indicate that the variance component for intercepts is weakly significantly difference from zero (*z-statistic*=1.71, *p*=0.043). This suggests that there could be additional variation in group mean energy levels that are not explained by these two factors. Therefore, it is expected that there may be other group-level factors that might explain the variance in group means. The variance components for slopes was estimated to be zero, and the component that represents the covariance between intercepts and slopes is small (-0.0146) and the null hypothesis that it is also zero cannot be rejected (*p*=0.786).

Table 7.28: Covariance parameter estimates of fixed and random effects controlling for the number of occupants at the individual and group level

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	GROUP_CODE	1.556	0.9081	1.710	0.0433
UN(2,1)	GROUP_CODE	-0.01459	0.05381	-0.2700	0.7863
UN(2,2)	GROUP_CODE	0	.	.	.
Residual		94.21	3.342	28.19	<.0001

In order to improve the model, several other individual-level and group-level variables could be used to help explain more of the variation in individual household energy use (fixed effects) and in group mean energy use (covariance parameters). Some group-based variables, including belonging to an urban area, a suburban area, did not however show any significantly different results.

There were two binomial individual-level variables tested to see if they could improve the model. The first was the variable for dwelling type (house or flat) and for building age (pre-1946 and post-1945). The first, dwelling type, showed that there were two models with significantly different intercepts, but that there was no interaction between dwelling type and household size nor between individual household size and group mean household size.

Table 7.29: Solution for fixed effects controlling for the number of occupants at the individual and group level, and for the number of rooms

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	54.56	0.8433	18	64.68	<.0001
group_mean	0.9444	0.2061	18	4.580	0.0002
house	1.863	0.8441	1585	2.210	0.0275
hsize	1.168	0.1071	1585	10.90	<.0001
group_mean*hsize	-0.01071	0.01516	1585	-0.7100	0.4800
house* hsize	-0.1702	0.1106	1585	-1.540	0.1240

However, for building age, not only were there two models for pre- and post-war housing that had significantly different intercepts, but in these models the slopes of individual household size differs depending on their group mean household size, which did not occur in the simpler multilevel model nor in the model that included dwelling type. In addition, the interaction between individual household size and building age tells the researcher that the slopes for individual household size are significantly different for the two building age categories. A test was made to see if the additional predictor variable of dwelling type could have been included, but the interaction between individual household size and their group mean household size was not significantly different from zero.

Table 7.30: Solution for fixed effects controlling for the number of occupants at the individual and group level, and for the age of the dwelling

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	55.40	0.4716	18	117.47	<.0001
group_mean	0.9807	0.2391	18	4.100	0.0007
prewar	1.402	0.5220	1583	2.680	0.0073
hsize	1.163	0.04342	1583	26.78	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
group_mean*hsize	-0.04600	0.02267	1583	-2.030	0.0426
prewar*hsize	-0.2330	0.06178	1583	-3.770	0.0002

Therefore the model with the interactions with and including the dummy variable for pre- and post-war housing can be written as:

$$\sqrt{e_{ij}} = \beta_{0j} + \beta_{1j}(x_{ij} - \bar{x_j}) + \beta_{2j}p_i + \varepsilon_{ij} \quad (94)$$

Where p_i is the dummy variable for pre- (1) and post- (0) war housing.

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\bar{x}_j + u_{0j}, \quad (95)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\bar{x}_j + u_{1j}, \quad (96)$$

and

$$\beta_{2j} = \gamma_{20} + \gamma_{21}\bar{x}_j + u_{2j}. \quad (97)$$

This results in the equation

$$\sqrt{e_{ij}} = \gamma_{00} + \gamma_{01}\bar{x}_j + u_{0j} + \left(\gamma_{10} + \gamma_{11}\bar{x}_j + u_{1j}\right)(x_{ij} - \bar{x_j}) + \left(\gamma_{20} + \gamma_{21}\bar{x}_j + u_{2j}\right)p_i + \varepsilon_{ij} = \left[\gamma_{00} + \gamma_{01}\bar{x}_j + \gamma_{10}(x_{ij} - \bar{x_j}) + \gamma_{11}\bar{x}_j(x_{ij} - \bar{x_j}) + \gamma_{20}p_i + \gamma_{21}\bar{x}_j p_i\right] + \left[u_{0j} + u_{1j}(x_{ij} - \bar{x_j}) + u_{2j}p_i + \varepsilon_{ij}\right] \quad (98)$$

Again, the covariance parameters suggest that the variance component for intercepts remains significantly different from zero. As before, the null hypothesis is not rejected that slopes do not differ between groups, nor that there is no covariance between slopes and intercepts.

Table 7.31: Covariance parameter estimates of fixed and random effects controlling for the number of occupants at the individual and group level, and for the age of the dwelling

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	GROUP_CODE	1.749	0.9952	1.760	0.0395
UN(2,1)	GROUP_CODE	0.01783	0.05918	0.3000	0.7632
UN(2,2)	GROUP_CODE	0	.	.	.
Residual		93.11	3.307	28.16	<.0001

This implies that only random intercepts for groups u_{0j}, and not random slopes, are appropriate for the fitting of this model. The model is then re-run without allowing random slopes, leading to significant group variance τ_{00} and individual residual σ^2:

Table 7.32: Covariance parameter estimates of fixed effects controlling for the number of occupants at the individual and group level, and for the age of the dwelling (intercept only)

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	GROUP_CODE	1.750	0.9939	1.760	0.0391
Residual		93.12	3.3070	28.16	<.0001

and the following fixed effects for the intercept, or overall mean, with the following parameters for the group mean household size, individual household size, their interaction, the building age, and the interaction of building age and individual household size:

Table 7.33: Solution for fixed effects controlling for the number of occupants at the individual and group level, and for the age of the dwelling, intercept only, decennial option

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	55.41	0.4646	18	119.26	<.0001
group_mean	1.116	0.1943	18	5.740	<.0001
prewar	1.308	0.5138	1585	2.550	0.0110
hsize	1.162	0.04337	1585	26.80	<.0001
group_mean*hsize	-0.03105	0.01522	1585	-2.040	0.0415
prewar*hsize	-0.2403	0.06134	1585	-3.920	<.0001

This results in two models, one for prewar (to be precise, pre-1946) and one for post-war (post-1945) housing. Substituting in the number 1 for pre-war and 0 for post-war:

Pre-war:

$$\sqrt{e_{ij}} = (55.414 + 1.308) + 1.116\bar{x}_j + (1.162 - 0.240)(x_{ij} - \bar{x}_j) - 0.0311\bar{x}_j(x_{ij} - \bar{x}_j) \quad (99)$$

Post-war:

$$\sqrt{e_{ij}} = 55.414 + 1.116\bar{x}_j + 1.162(x_{ij} - \bar{x}_j) - 0.0311\bar{x}_j(x_{ij} - \bar{x}_j) \quad (100)$$

Because group mean household size $\bar{x}_j$ is centred around the grand mean for England, and $(x_{ij} - \bar{x}_j)$ represents individual household size centred around its group mean, the parameter estimates above can be interpreted. The average energy use for pre-war housing is 56.722^2 = 3217 kWh per annum; in post-war housing this is 55.414^2 = 3071 kWh per annum. Both group mean household size and individual household size are associated with non-heating end-use energy in both pre- and post-war housing, although the magnitude of the effect is different. As pre-war housing gets larger, energy use grows at approximately 20 percent of the rate of post-war housing. Finally, there is a small interaction between individual and group mean household sizes.

7.4.8 Running the multilevel model with the estimation of 2008 energy use of homes in the 1996 English House Condition Survey (annual option)

Rerunning the same model for the annual option for 2008, similar results were found for the fixed effects in the random intercept only model:

Table 7.34: Solution for fixed effects controlling for the number of occupants at the individual and group level, and for the age of the dwelling, intercept only, annual option

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	55.44	0.4806	18	115.37	<.0001
group_mean	1.041	0.2014	18	5.170	<.0001
prewar	1.253	0.5204	1525	2.410	0.0161
hsize	1.158	0.04421	1525	26.20	<.0001
group_mean*hsize	-0.02826	0.01540	1525	-1.840	0.0666
prewar*hsize	-0.2418	0.06219	1525	-3.890	0.0001

Again, this results in two models, one for prewar (pre-1946) and one for post-war (post-1945) housing, substituting in the number 1 for pre-war and 0 for post-war:

Pre-war:

$$\sqrt{e_{ij}} = (55.442 + 1.253) + 1.041\bar{x}_j + (1.158 - 0.242)(x_{ij} - \bar{x_j}) - 0.0283\bar{x}_j(x_{ij} - \bar{x_j}) \quad (101)$$

Post-war:

$$\sqrt{e_{ij}} = 55.442 + 1.041\bar{x}_j + 1.158(x_{ij} - \bar{x_j}) - 0.0283\bar{x}_j(x_{ij} - \bar{x_j}) \quad (102)$$

In comparison to 1996, there was a higher overall mean energy use level in 2008, and differences in group mean household size and individual household size had slightly less impact on the amount of energy used in 2008 compared to 1996. This can be expected as household size in 1996 is an imperfect predictor of household size in 2008.

7.4.9 Comparison of Models using the Akaike information criterion (decennial option)

Using the Akaike information criterion (AIC), the models using the ONS area classifications at the group level were compared against each other:

Table 7.35: Comparison of Models using the Akaike information criterion, decennial option

Model	AIC (decennial)	AIC (annual)
Unconditional means ("empty") model	12830	12340
Group-level predictors only	12830	12340
Individual-level predictors only	11930	11460
Both group and individual-level predictors (random intercepts and slopes)	11900	11430
Both group and individual-level predictors (random intercepts)	**11890**	**11430**

This shows that the multilevel model that includes both the group and individual-level predictors but only allows intercepts to vary instead of both slopes and intercepts is the most reliable model for predicting non-heating end-use energy.

7.4.10 Verification of the annual model using actual data from 2008

Verification of the multilevel model is difficult because building age data is held by the Valuation Office Agency (VOA) as part of the private information that they keep to place dwellings into council tax bands (Valuation Office Agency, 2011). The VOA does provide aggregated totals for 15 dwelling age bands per local authority. Dwelling age changes drastically between area classifications (e.g. between an inner city urban and a rural location within the same local authority). Therefore, this is not a reliable statistic to use as part of the cross-tabulation of household size as in the single-level model earlier in this section.

7.5 Discussion and Conclusions

Two main conclusions were drawn from the two approaches to the single-level and the multilevel model. The first conclusion is that the single-level, annual model is the more reliable of the two models of non-heating end-use energy because it can be more readily verified. The second conclusion is that the multilevel model has a higher level of sophistication with similar information requirements – the size of the household as measured by the number of rooms and the number of occupants – but with an additional requirement of knowing the rough building age (pre- and post-war).

The validation of the single-level model has three issues to consider. First, the nature of the square-root transformation of energy use increases the impact of any extremes, possibly increasing the likelihood of an overestimate in the model at an individual level. Second, the value for electricity use as a proxy is likely an overestimate of non-heating end-use energy as there is no guarantee that

electricity use for heating has been eliminated altogether, increasing the likelihood of overreported energy use in aggregate. Third, there is an upper limit to household size available from the census statistics, but no cap on the energy use predicted in the model. The model estimate is very accurate overall with a slight tendency to overestimate energy use.

There were some efforts to overcome these tendencies in the model. First, sizes of homes 2 standard deviations above and below the mean in the housing survey dataset were labelled as high leverage points and eliminated. Second, an average household size value for “above-category” categories (e.g. 6+ rooms in a household) in census data was used to estimate the size of homes in that part of the residential sector. There was no basis used to attempt to alter the reported electricity use of a LLSOA.

Comparing the intercepts and slopes of the single-level and multilevel models requires some additional interpretation. The intercept of the single-level model is the non-heating end-use energy when the household size is zero. The intercept of the multilevel model is the average non-heating end-use energy across the residential sector. This is because the predictor values of household size have been centred on the grand mean household size for England, and so individual household sizes are centred around their ONS area classification group mean. The slopes, or parameters, in the each of the models both represent the influence of additional household size at the same scale.

The stable version of the multilevel model requires for its validation building age data that, unfortunately, is not publicly available in aggregated form across England and cross-tabulated with household sizes. However, the model does have the appropriate predictive power that can be generated with ease for England given appropriate access to the data. The multilevel model does reveal the impact of group and neighbourhood membership on individual household end-use energy with almost half of the explanatory power of household size contained within the group and not the individual household, and this is part of the discussion of the results and further work in Chapter 8.

Chapter 8 - Discussion and further work

8.1 Introduction

This chapter discusses the implications of the results as reported in Chapter 7 on domestic energy modelling, from the models that were explored in full and the models that were abandoned during the research process. The first section discusses the implications of the single-level model in comparison to the one currently in operation in England as part of the Standard Assessment Procedure and the bottom-up housing stock model BREHOMES. The second section discusses the implications of adopting a multilevel model, allowing the area and building type to affect the predicted energy use of existing buildings in order to build a more accurate model but maintaining the use of the interaction term. The third section asks if the two abandoned models or further experimental methods are realistic candidates for modelling non-heating end-use energy in the future. The final section discusses how the measurement of this type of energy use fits in with the entire exercise of quantifying sustainable living in the 21st Century.

8.2 Implications of adopting the single-level model

8.2.1 Introduction

The single-level model that is proposed in the results section is a simple regression that uses household size as a predictor and non-heating end-use energy as the outcome. There are positive implications for the use of this interaction term, as validation using basic information from the census and aggregate energy data that has only lately become available is now possible. In selecting a statistical model, the goodness of fit needs to accompanied with access to data that is reliable in the medium to long-term. However, this verification assumes electricity consumption is the same as non-heating end-use energy demand in LLSOAs that overwhelmingly use domestic central heating. The outcome variable had a positive skew that required transformation and exclusion of outliers in order to normalise the data for use in linear regression. This has an effect on the building of a bottom-up domestic energy stock model for these end-uses. These measures affect the makeup of any model of non-heating end-use energy in the residential sector made from this data. As electric heating is not totally isolated, models using electricity meter data will overestimate non-heating end-uses. Homes with extreme energy consumption, both large and small (but overwhelmingly large), are not included in the data. Are these large consumers driving overall increases in consumption? Using traditional regression techniques, researchers cannot accurately model these

households, which essentially means that they are modelled *sui generis.* These are interesting talking points to be explored around bottom-up domestic energy stock models.

8.2.2 Change from floorspace to rooms as measure of physical household size

The most positive benefit from the adoption of this single-level model for the estimation of non-heating end-use energy is the technique for validation of the results using actual energy use consumption data. Previous attempts to validate the individual household-level model used in a bottom-up housing stock model for England were done against a second model of appliance ownership that assumed usage of each category of appliance (Shorrock and Dunster, 1997, Environmental Change Institute, 1995).

The current model at the individual household, level, called BREDEM or SAP, measures household size in usable floorspace, which is only available for homes that participate in housing surveys and possibly available by searching past planning and building permission applications dependent on the scale of the project, the number of units, and the archives kept by each local planning authority. The number of habitable rooms, however, is a measure that has been kept in the United Kingdom Census, the regional planning of housing, and in applications for building control for the past century (Martin, 2001, Abercrombie and Forshaw, 1943). There is a consistency in the definition of a habitable room, and this information is likely to continue in the future. This suggests that the number of rooms will continue to be the most widely collected measure of physical household size. Culturally, the English do not buy or rent housing using as a primary basis for their decision the amount of floorspace as measured in square metres, even though this information may be available, but the number of rooms or bedrooms. Anecdotally, if one asks a native of other European countries, they will know exactly what the number of square metres means as a measure of physical household size – for the English, they will be much more uncertain until the number of bedrooms or rooms is stated. This cultural predisposition makes the interaction of rooms with the number of occupants more relevant to the measurement of household size than the interaction of usable floorspace with occupants, even though the former is a less detailed measurement than the later.

Also available for use by academics and researchers are the cross-tabulations of occupants with different household sizes to investigate in more detail the impact of the interaction between occupants and the number of rooms (ESRC Census Programme, 2006). The validation procedure can then model each of the household sizes in turn (2 room, 1 person or 3 room, 3 person, etc.) multiplied by the number of households in each size category instead of using an interaction term built out of the mean for that census area.

There may be an opportunity in the future to cross-reference the size of anonymised households with their metered electricity for the purpose of refining models of non-heating end-use energy. This thesis has shown the potential for a model for this type of energy use considered separately from the heat flux equations designed for measuring heating demand. The use of the number of rooms and the number of occupants in a household can also encourage more data to be made available from households whilst protecting their privacy. This can make it easy for community organisations, energy cooperatives, local authorities, or even commercial energy companies to more easily share electricity meter data (of houses that do not use electricity for their heating) to help refine domestic energy stock models of non-heating end-uses.

8.2.3 Aggregate data – ordinary electricity use as a proxy

This model is validated against the annual metered electricity use of Lower Layer Super Output Areas (LLSOAs) for England as provided by the Department for Energy and Climate Change, but electricity use is an imperfect proxy for non-heating end-use energy. This is a drawback of the model because there is no guarantee that equating central heating with not using electricity as the heating fuel is equally true across all areas. However, heating systems powered by electricity typically increase the amount of electricity used in a household six-fold, so great care must be taken to eliminate as many of these homes as possible. Another drawback is that if the heat fuel of the housing stock is changed to electricity to allow supplies from on-site microgeneration such as photovoltaics, verification becomes all too difficult and many become impossible without extensive sub-metering in domestic households that is not foreseen.

The Digest of UK Energy Statistics (Department of Energy and Climate Change, 2009a) shows that the gasification of central heating is nearly complete within the country, so the equation of central heating systems with natural gas as a fuel is not an outlandish assumption. However, the energy required by any home that requires electricity as the fuel for its heating system can easily skew upwards the reported electricity use of each LLSOA. This will not necessarily show up in the electricity figures provided by the Department for Energy and Climate Change in their small area statistics. The verification of the model uses only ordinary rate electricity meters, as it assumes that any electric storage heating – that is, a water tank heated by electricity – will be run at night using reduced rates offered in the UK and known as Economy7 after the seven overnight hours with preferential rates. These meters and the electricity used by them are not included in the verification of the single level model.

This situation will be muddled in the future as the “gasification” of housing is predicted to be reversed to bring more electric heating on stream using renewable sources that generate electricity

to supply heating (Boardman, 2005). However, electricity for different end-uses will likely not be sub-metered in homes when this occurs. The on-site generation of electricity can be measured, but this may or may not be directed towards a particular set of end-uses. Unless sub-metering of heating systems is made a part of the changeover from natural gas, our understanding of user behaviour and the connection between household size, built form, and energy use of all end-uses will deteriorate since verification of any model will become more difficult as this conversion process goes forward.

I argue that this spectre of an inability to validate any amount of delivered end-use energy in a household makes getting the right model today - when these end-uses are supplied from different fuels - all the more important. This is shown by this thesis's investigation of the history of the current BREDEM/SAP algorithms built from data on the heat flux, or the difference between indoor and outdoor air temperature of a building. For non-heating end-uses, direct validation of their contribution to the indoor temperature was impossible as it was the heat given off through the operation of an unknown amount of appliances and their cabling. This situation may occur again through the measurement of electricity use, as heat demand from electricity becomes more prevalent in the future. Finally, the use of secondary heating systems in poorly insulated dwellings can alter the total electricity demand when verifying against aggregate data. SAP2009 currently estimates that around 10% of all heat demand comes from secondary heating systems if a gas boiler is being used. In centrally heated homes, this could, if all these secondary heating sources are electric (which they are not), result in an increase of electricity demand of as much as 25 percent over that of lighting, appliances, and cooking in the UK. Fortunately, recent observational data places secondary heating demand to be around 3.6% of total delivered energy for space heating (Energy Saving Trust, 2009b).

8.2.4 Building a bottom-up domestic energy model with transformed data

This bottom-up domestic energy model was built from energy meter data that was transformed in order to meet the parametric tests for linear regression modelling. Firstly, this single-level model acknowledges that the main source of domestic energy meter data, the 1996 English House Condition Survey, was a stratified and not a simple random sample. The square root transformation contrasts with the natural logarithmic transformation used in SAP2009, which was the first time transformation of the data was done to create a model of domestic non-heating end-use energy. It should be noted that only the dependent variable was transformed in this work, where as both the dependent and independent variables were transformed in SAP, and that SAP allows for calculated, and not actual numbers of occupants as a function of floorspace (BRE, 2010). Finally, the relationship of the modelled energy of a household and the total amount of energy of the domestic sector needs

to be interpreted with caution using the single-level model, as transformation and exclusion of outliers and high leverage points changes this relationship.

Housing surveys that can accurately reflect the makeup of the national population are expensive and time-consuming, and therefore are only conducted by governments. As explained earlier in this thesis, even with these resources, a stratified sample was necessary to have enough datapoints for different types of housing, notably those which were part of the social housing sector. As this model limited itself to houses without electric heating, the makeup of the sample was altered and modelling proceeded as if the sample was random. This process ignored the fact that the selection of just homes that did not use electric heating was, in itself, a new stratum put into the sample. Electric heating is more likely to occur in homes at the extreme ends of building age – period buildings that are difficult to convert to central heating, and modern homes on brownfield sites without a previous natural gas connection where the developer did not invest in and the local planning authority did not insist on a connector to supply new housing. High-rise housing is also overrepresented in the use of electricity for heating. This leads to a weakness in the model by the under-inclusion of urban housing in city centres and rural housing and the overrepresentation of suburban and mid-urban neighbourhoods. However, this is not an entirely dire predicament, as housing with its lowest floor level of five storeys or above comprise less than 1% of the total housing stock in England (Office for National Statistics, 2005c). On balance, the treatment of the housing survey data as a simple random sample was the right choice. In the future, a fuel sub-sample of housing surveys, such as the one that is about to be commenced off the back of the 2011 English Housing Survey must ensure that the type of fuel is a stratum itself as well as tenure, housing type, and building age.

The decision to transform the dependent variable of non-heating end-use energy by taking the square root instead of a natural logarithm is a major deviation from the current BREDEM/SAP model. The advantage of a natural logarithm when it is back-transformed to $y = ax^c$ is, if c is less than 1 but greater than 0, as it is calculated in SAP2009, a non-linear model is produced of diminishing energy outcomes as household size grows. If there are very large houses, they are predicted to use energy at a similar level to other large houses because the model predicts a theoretical maximum amount of non-heating end-use energy in a household. Instead, when a square root transformation is applied to the data, it is back-transformed to $y = (a + bx)^2$ which accelerates the increase of predicted energy use of homes as they get larger, and decreases the amount of energy predicted for low-to-middle household sizes. This functional form can behave in possibly perverse manners for very large homes (with an interaction term greater than, say, 20, which would represent a family of four in a five-roomed, or three-bedroom house). It may be necessary to estimate these homes

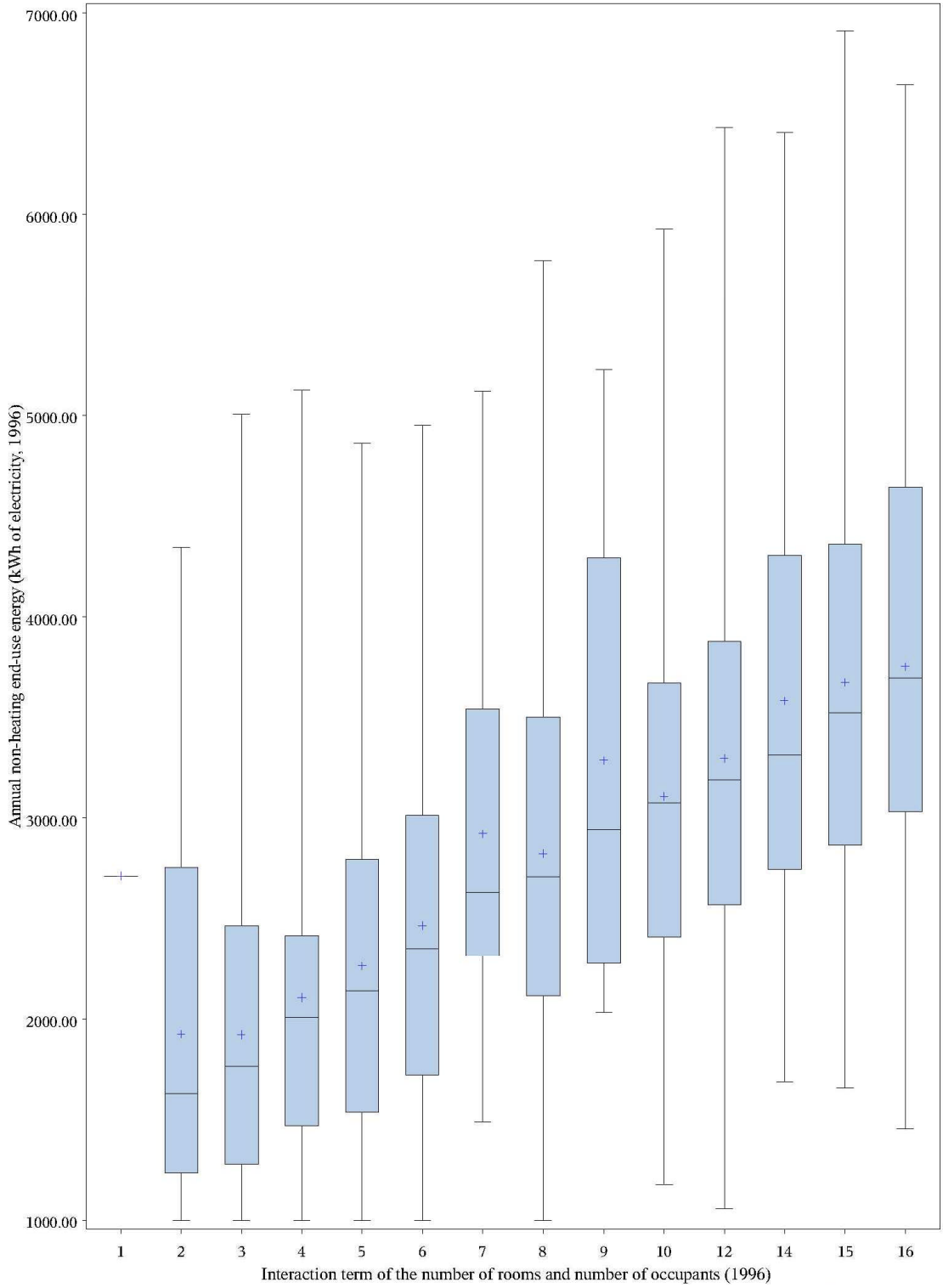
outside the scope of these models, or at the very least without adding in any group-level interactions. Table 8.1 illustrates this by comparing the single-level model in this thesis as applied to a household with SAP2009 without correction factors using given numbers of bedrooms and occupants.

Table 8.1: Example households comparing this thesis with SAP2009

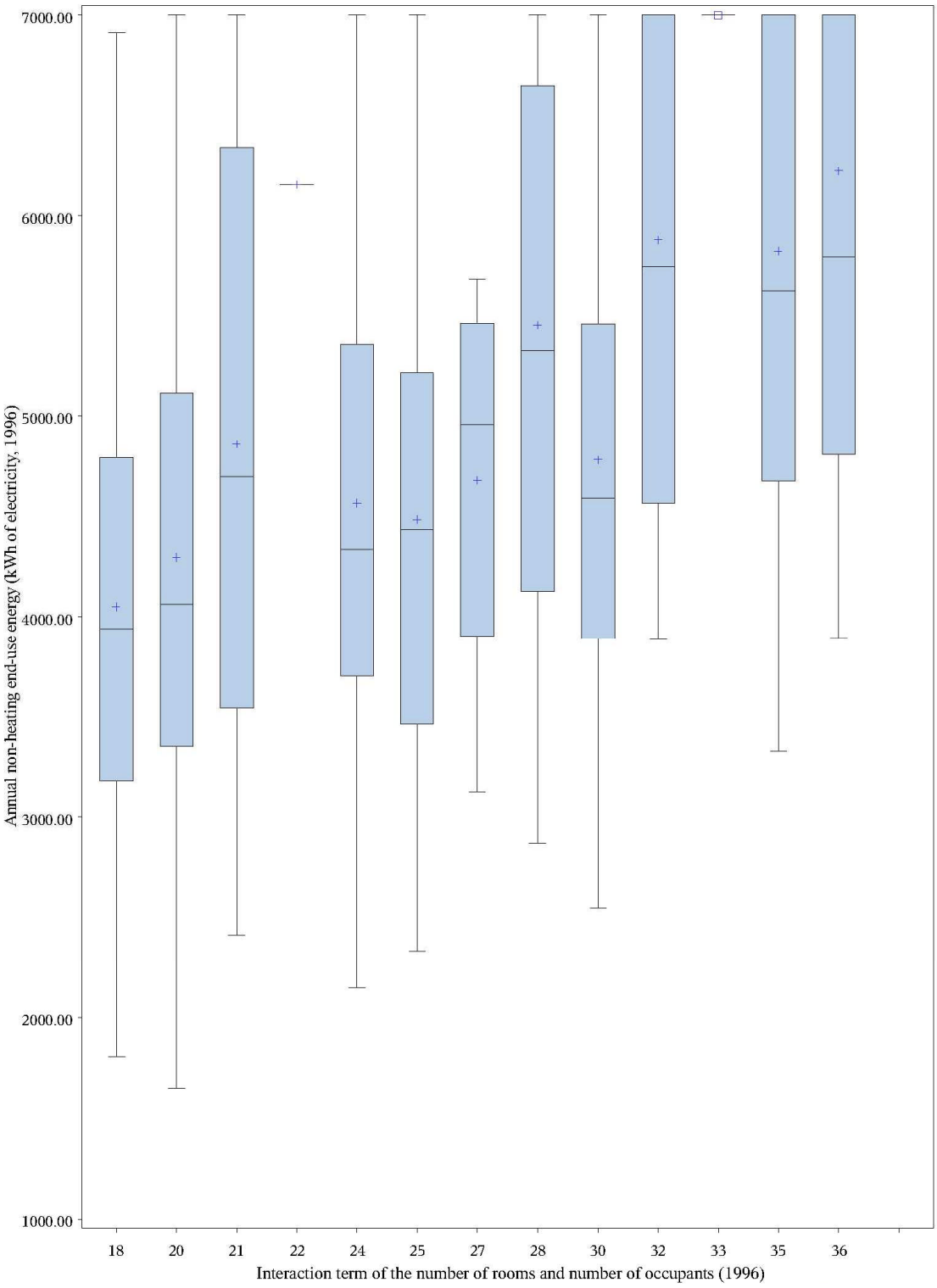
	A	B	C	D	E	F
Bedrooms	1	1	2	3	3	4
Occupants	1	2	3	3	4	5
Estimated habitable rooms	3	4	5	6	7	9
Estimated floorspace (m2)	50	60	80	100	120	160
Interaction (occupants x rooms)	3	8	15	18	28	45
Interaction (occupants x floorspace)	50	120	240	300	480	800
Thesis (kwh/year)	2099	2608	3413	3791	5194	8084
SAP2009 (kwh/year)	1691	2556	3543	3936	4913	6250

Qualitatively, as showed in Figure 8.1, there does not appear to be a definitive direction (curving up or down) in the data – it appears almost linear. However, the square root transformation does have evenly distributed residuals, as shown in Figure 8.2.

The amount of people per room is expected to continue to decrease in the future, so there will be an impact on the kind of homes that are in the middle – there is a smaller number of people using the same space, but with the same basic infrastructure of appliances. This may necessitate a model in the future based either on an interaction term that weights the physical and occupant sizes differently if our understanding of their contribution increases in the future, or the development of models that incorporate a minimum for their estimation of household non-heating end-use energy.



Part (i) of Figure 8.1



Part (ii) of Figure 8.1

Figure 8.1: Boxplots of the annual non-heating energy use by the interaction term from the English House Condition Survey

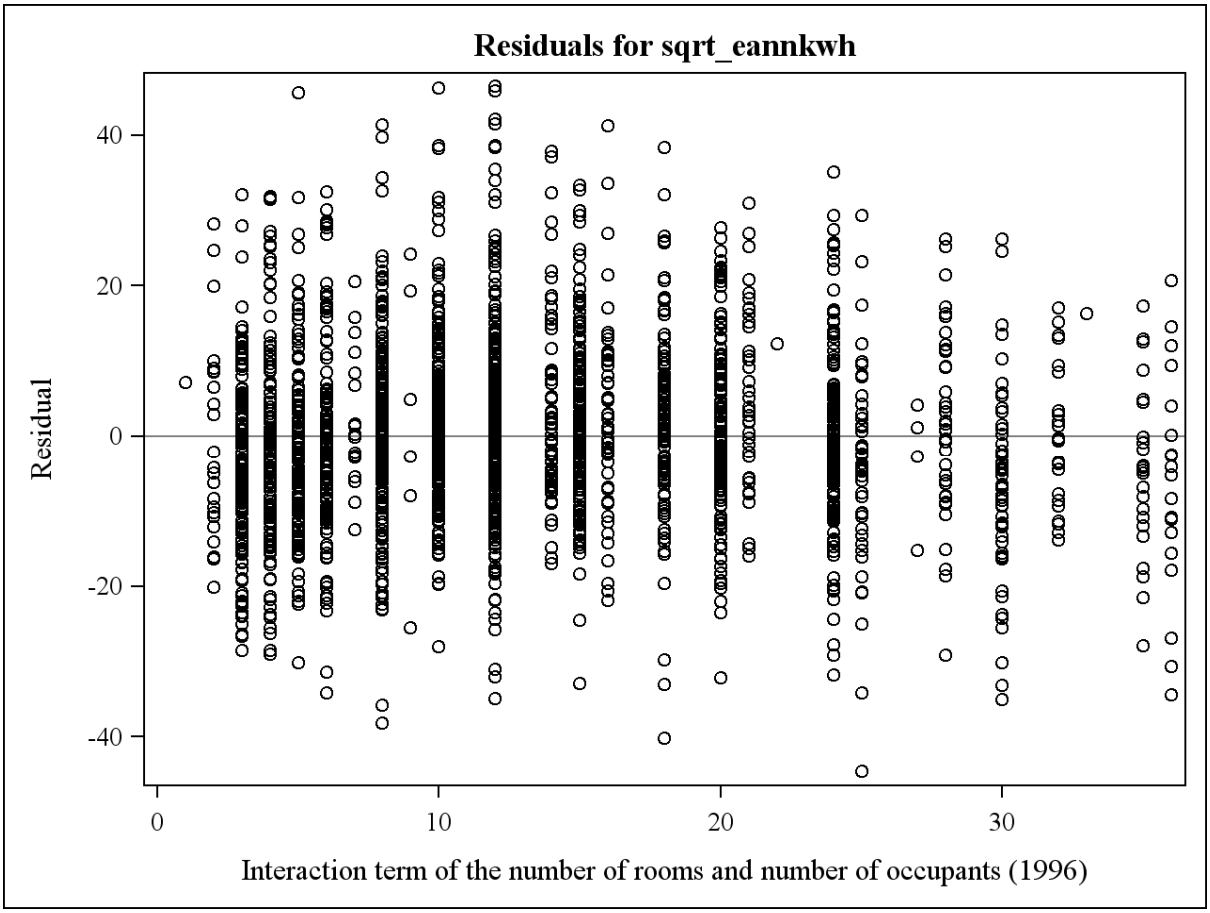


Figure 8.2: Residuals for energy use after transformation across all values of the interaction term

This development of a minimum level would be a positive reflection on the data that researchers currently have on non-heating end-use energy consumption. The combination of no minimum and the removal of outliers is likely leading to an overestimation of the energy consumption of smaller sized households, which can overly promote energy conservation measures being targeted at smaller-sized (and more likely to be lower-income) households. The converse problem is that as high-consuming outliers are excluded from the construction of the model, they are not subsequently studied in detail to find why their energy use is so exceptional.

The exclusion of outliers as a requirement of creating a reliable household model built from a normally distributed dependent variable, albeit transformed, also changes the nature of a domestic energy stock model. The total predicted energy use of an entire population, such as a census area or nation, is modelled from a sample, making this total amount of energy use not a "population total," but a "sample total." Since there was such a large positive skew in the level of non-heating end-use energy found in households, there are two conclusions that can be drawn. The first is the assumption made in this thesis: that the outliers represent households that cannot be part of the population of households that do not use electricity for heating. The second possibility is that they

are, and use electricity in a particularly prodigious way that is not encompassed within the household model.

The imperfect relationship between the "sample total" and the "population total" has not been fully explored by delving into study of these outliers in more detail. With energy use, this matters, both from the perspective of meeting national obligations for climate change, and for the planning of future electricity infrastructure that will not be able to operate with the same amount of extra available generation capacity to satisfy peak demand in the future.

8.2.5 Conclusions

Bottom-up domestic energy stock models are driven entirely by their relationship with the modelled households, and this will not change in the future. As it treats every household equally, the single-level regression model of household non-heating end-use energy might treat different parts of the spectrum of household sizes unfairly. As shown in Figure 8.3, the current logarithmic transformation likely over-predicts average-sized housing, and the proposed square root transformation likely over-predicts large size housing.

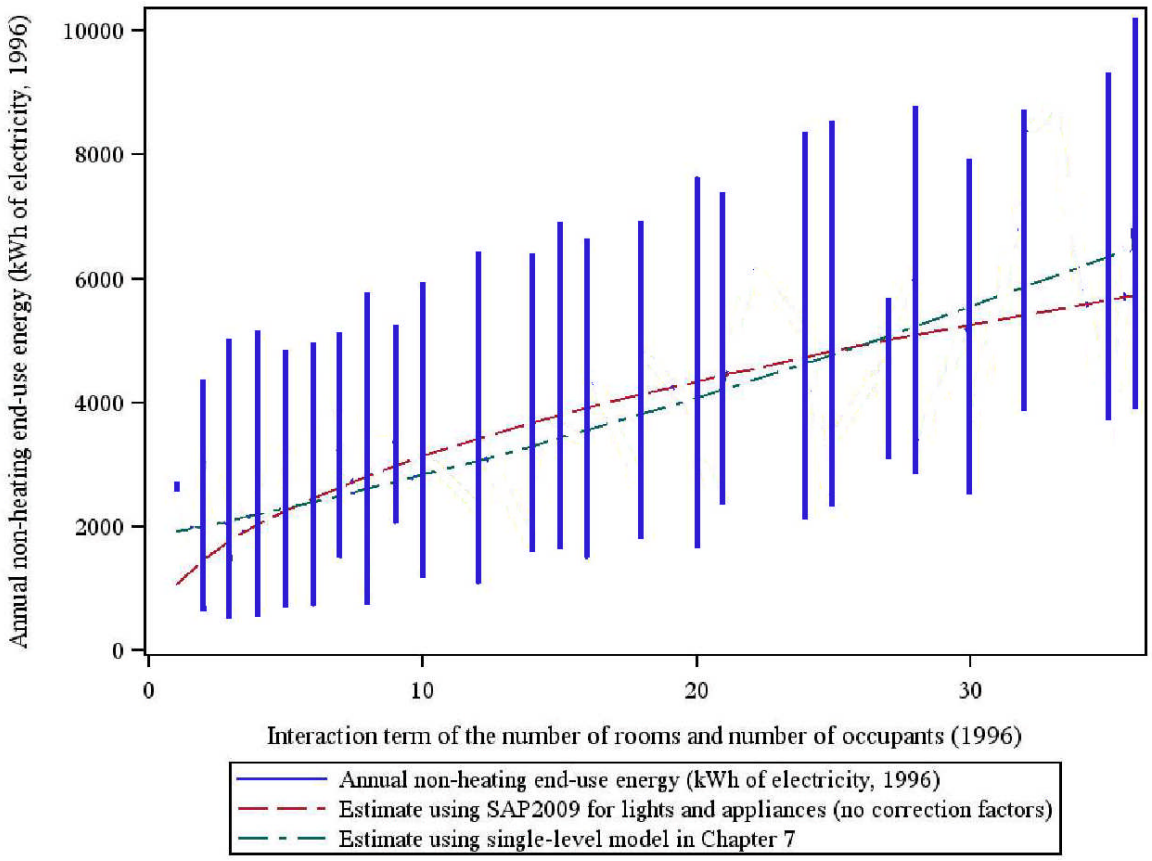


Figure 8.3: Comparison of SAP2009 and single-level model presented in Chapter 7 against ranges of electricity use in homes without electric heating in the 1996 EHCS

With the advent of new datasets that enable these models to be verified against aggregate electricity use data, community groups and local authority planners may begin to verify the national stock model against just their neighbourhood or town. The resulting discrepancies will lead to local scrutiny of these models and pointed questioning from local elected officials as to their applicability to their community.

These issues lead to the question posed by the multilevel model – it is a proper approach to treat each individual household differently based on the group, or in the case of this thesis, the area classification, to which they belong? And what criteria are appropriate? They could be socioeconomic, geographical, or both, which was the compromise position taken in the analysis earlier in this work. The multilevel approach might well introduce more overt inconsistencies and perceived unfairness, but the counter argument is that the single-level model does likewise, but covertly, as few non-specialists understand the implications of transforming data to use linear regression techniques.

8.3 Multi-level model findings

8.3.1 Introduction

A multi-level model was proposed as a viable alternative to the single-level model, but one that was more difficult to verify as the data available is incomplete. The main effect of the innovation of the multilevel model was simultaneously use individual household size and average household size of different area classifications to predict energy consumption. Although the socioeconomic and built environment characteristics of area classifications do not directly estimate energy consumption in place of mean household size, this method should be seen as a small step towards using the setting of the home as a predictor of energy use. This would be useful as a way of rescaling energy models to take account of likely socioeconomic statuses and built environment constraints that could be affecting consumption in the individual household.

The interaction between the two levels without any additional predictors was not significantly different from zero, but building age (pre- or post-war) as an additional predictor stabilised the model, and the interaction between the two levels became statistically significant. The model minimises the number of predictors and maintains the simplicity of the interaction term of numbers of rooms and occupants found in the single-level model. This has the drawback, however, of a model

that is currently difficult to verify unless aggregate data at the LLSOA level for building age is released for England.

8.3.2 Group-level effects

Using the mean household size of an area classification as a predictor of the non-heating end-use energy of a model is likely to be seen as a controversial step in the development of domestic energy stock models. The nature of these classifications and their limitations as clusters of characteristics and not of territory should be understood. Also, many may assume that the mean energy level of an area should be the predictor, but the nature of a multilevel model directs the researcher to use the mean of the same independent variable(s) used at the individual level to measure between-group and within-group variance in energy use. This may seem counter-intuitive for some decision-makers. By using a variable from an area classification of a neighbourhood, two new considerations are now in play. First, there is the question if building regulation standards in the form of the target emissions rate are permitted to vary not just according to dwelling specifications, but also to neighbourhood characteristics. Second, when a person measures their use non-heating end-use energy, this type of model would tell an individual that they are only partly responsible for their energy use within the homes, and the rest is due to the homophily involved when choosing an area to call home. These considerations lead not just to an interpretation of these end-uses as explained as a physical construct, but also explained as a social construct of domestic energy stock modelling.

Privacy rules and undertakings made to the participants of the 1996 English House Condition Survey mean that location data below the regional level cannot be released by the holders of the data, the Department for Communities and Local Government in England. The nature of the area is given by the area classification for the LLSOA within which the surveyed home is located. Area classifications are created by the Office of National Statistics using clustering algorithms (Vickers, 2006). The *k-means* method attempts to associate similar areas by reducing the numerical difference between areas using a range of characteristics from dwellings, the surrounding built environment and socioeconomic data of the occupants of dwellings in each area, but also has the disadvantage of accentuating the differences between area classification groups, especially when two groups are spatially adjacent to one another. As overall statistics for different area classifications are calculated as mean values, they represent a wider range of these characteristics of different neighbourhoods. For example, a classification area may be typified by a higher proportion of flats, but there could be neighbourhoods near the bottom of this range of flats that are closer to nearby neighbourhoods that have a high level of flats within a different classification typified by a more moderate level of flats. If the location data was made available, this type of statistical feature would be avoided, but the data is unlikely to be given to researchers in the near future (McIntyre, 2011).

The use of a common interaction term that combines two independent variables can be confusing for politicians or the general population without a strong mathematical or statistical background. Intuitively, many would wish to use the mean annual household non-heating end-use energy as the predictor variable at the group level instead of the mean of the interaction term. Statistically, this type of relationship is not permitted as it is using a dependent variable as an independent variable at a different level. Without using the same interaction term of the two independent variables, effect sizes would become only relative and not absolute, and lose their power to predict energy use in annual kilowatt-hours in a bottom-up domestic energy stock model made up of households.

Within the legal bounds of building regulations, many would argue that the group dynamic in the area should not affect resulting target emissions rate that is calculated from the design parameters of the dwelling. However, the group dynamic in other aspects does impact on the obtaining of planning permission. In the case of England, national planning guidance requires that any development, including individual dwellings, must take account of their impact on the design quality of a neighbourhood, be appropriate for the housing market to achieve mixed communities, and be provided in a suitable location (Department for Communities and Local Government, 2010b). Standards in force for building regulations do not consider neighbourhoods, but rating systems such as BREEAM, LEED-H, and SB-Method (Prior and Williams, 2008, US Green Building Council, 2008, International Initiatives for a Sustainable Built Environment, 2010) that are in use in much of the developing world give credit not only to the energy consumption predicted for the building, but of the benefits of its location and relationship of its neighbourhood with the rest of the area, be it urban, suburban, or rural.

Table 8.2: Ranks of mean scores for ONS area classification groups

2001 Area Classification LLSOA Group Name	Interaction Term Score	Interaction Rank	Density Rank	Social Rented Rank	Dwelling Type Flat Rank	Public Transport Rank	Routine Employment Rank
Farming and Forestry	16.4588	1	20	17	19	20	14
Urban Commuter	15.9137	2	16	20	20	16	16
Affluent Urban Commuter	15.4244	3	15	19	12	11	17
Multicultural Urban	15.0505	4	3	10	10	5	9
Rural Economies	14.6953	5	18	15	16	19	15
Young Urban Families	14.5048	6	13	16	18	15	8
Countryside Communities	13.7313	7	19	8	17	18	6
Well off Mature Households	12.9879	8	14	18	15	14	11
Mature Urban Households	12.5846	9	17	7	14	17	5
Suburbia	12.4406	10	9	9	9	7	10
Blue Collar Urban Families	12.205	11	11	3	13	9	1
Urban Terracing	11.742	12	7	13	11	10	3
Mature City Professionals	11.4434	13	6	14	4	4	18
Small Town Communities	11.2942	14	12	5	8	12	4
Multicultural Suburbia	11.1423	15	5	4	6	3	7
Struggling Urban Families	10.5737	16	8	1	7	6	2
Educational Centres	10.5303	17	4	11	3	8	20
Resorts and Retirement	10.1558	18	10	12	5	13	12
Multicultural Inner City	9.5897	19	1	2	2	1	13
Young City Professionals	8.3682	20	2	6	1	2	19

Table 8.2 above lists the different area classifications by the rank of their mean interaction term score, or the number of rooms multiplied by the number of occupants. Some of the indicator variables were selected from the ONS 2001 Area Classification for LLSOAs including population density, prevalence of flats, social rented housing, flats, taking public transport, and routine or semi-routine occupations. The mean values for each of these variables were also ranked by their Area Classification Group. There were some expected results: places with large households sizes were low density, low in socially rented housing, and low in taking public transport. It was also not surprising that the two groups named "Multicultural Urban" and "Resorts and Retirement" were exceptions and behaved differently from other groups with similar household sizes. However, there was no apparent connection between areas with larger mean household sizes and the prevalence in that type of area of workers in routine or semi-routine occupations, which is often used as a proxy for measuring those in low-paid work and therefore the relative wealth of a type of area.

This is an interesting finding as the connection explored in this thesis relating to mean household size at the group level can partly explain energy use in individual households. This measure of household size does not attempt to measure attitudes, habits, or the ability to pay for the use of this electricity. Early studies indicated that lower income homes owned fewer electricity-consuming devices and used them less often (Energy Advisory Services, 1996). Recent studies have centred on the reduction of demand for heating, yet predicted that there would be no reduction in non-heating end-use energy in homes in fuel poverty and that they would continue to rise as quickly as homes that were not in fuel poverty (Ekins and Dresner, 2006). In addition, conditional demand analysis models that assumed the number of appliances and usage patterns based on socioeconomic patterns have largely fallen out of favour. Although there is considerable effort made to deduce the types of purchases of electronics, and to a lesser extent, appliances, in the homes, there is very little that can be said about the socioeconomic factors that influence choices to operate and maintain products after they have been selected, purchased, and taken into the home.

If the nature of an area has predictive power through the mean interaction term for household energy use, then there could be a moral hazard through the promotion of further fatalistic attitudes about one's own personal responsibility for energy consumption in the home. An argument that could emerge from the adoption of a multilevel modelling approach is that the findings can imply that once a household has "chosen" a neighbourhood, the type of area that

they live in accounts for almost half of the energy use, as the mean interaction term explains 45 per cent of the variance in annual household non-heating end-use energy. The model does not go to the level of the device, so consumer choices as a consequence of socioeconomic status likely found in that area classification are still uncertain. In addition, the exact location of the home is not included, so local energy saving initiatives cannot be taken into account that is changing the energy consumption of a committed group.

The model cannot measure the number of devices owned or the tendency of each individual occupant within a household to use them. There are too many poorly-understood behavioural variables in this type of approach. Studies of one's own personal contribution to energy use in the home depend on relating metered energy data at short time intervals (at the maximum half-hourly, which is the standard timescale for meter data recordings in the UK) with diary data, which are self-reported. The differences between reported and exhibited occupant behaviour using appliances, electronics, lighting, and cooking are poorly understood, with most studies centring on the behaviour of occupants' engagement with their space or water heating systems (Shipworth, 2010).

This model is not able to take account of community initiatives that reinforce increased energy efficiency and demand reduction (McMichael, 2009). More formal interactions between institutions and the local population should not be overlooked. Examples of these are information programmes on energy conservation run by local energy companies or the public sector as well as local planning policies enacted at the local or regional level.

These issues lead to the conclusion that the predictor at the group-level of mean household size represents not just physical construct of rooms and people, but is a reflection of a social construct built from the household itself and the type of area within which the household resides. Without using the same predictor variables of household size at the group level as are used at the individual level, the effect sizes would only be relative and not absolute measures of the differences in the use of metered energy consumption.

8.3.3 Effect of building age

The age of a dwelling as a predictor of non-heating end-use energy was somewhat unexpected, both in its significance, and for the lack of significance of individual and area-based factors such as settlement type (e.g. urban, suburban, rural) and dwelling type. The use of age as an additional predictor variable does enable significant predictive power for the model without altering the principle that the coefficients are interpreted as absolute and not relative increases in the number

of kilowatt-hours. Building age could be a proxy for a number of socioeconomic factors in the population that affect energy use. However, future work, using both social science and engineering methods, needs to be done to measure these connections between choice of housing, behaviour, and non-heating energy use.

The multilevel model does not have an interaction between the individual-level effect on energy use and the group-level effect on energy use until the dummy variable of building age of the dwelling, split into pre-war and post-war, is inserted into the algorithm. This is a binomial categorical predictor variable and not a continuous predictor variable, so it will not violate the setup of the model that allows interpretation of the coefficients as absolute, and not relative, effect sizes at both the group and the individual level.

The effect of a building belonging to an age profile of pre-war and post-war is a very interesting one to explore in future models that might predict behavioural factors from the socioeconomic profile of occupants. First, there may be a connection between the age of the occupants and the age of the buildings. Second, there could be a connection between the tendency to live in a modern house and living as a family unit with children involved with some additional effect on non-heating energy use. Third, there can be an addition knock-on effect of housing replacement in not just embodied energy, but in the day-to-day domestic energy use. Finally, the connection between the individual building type being pre-war and the entire area also being pre-war should also be explored, where an individual and group level effect can be separated in the energy use of buildings.

One of the more interesting results from the housing survey is that the other variable that enabled the interaction between the group-level mean household size and the individual-level household size to be significantly different from zero (effect size -0.031, $p<0.001$) is if the head of the household was over 60 (effect size 1.43, $p=0.0217$). This can be interpreted as meaning that if there are elderly residents, the household size effect on non-heating energy use is slightly less that other households. Not only is this an indication of the presence of more elderly people, but conversely there is also an absence of young people in those households. In the fuel sub-sample of the 1996 English House Condition Survey, households where the head of the household was over 60 had a mean value for the number of children as 0.022 (sd=.201), and households where the head of the household was under 60 had a mean of 0.928 (sd=1.12). However, this is a mean number of children, and there would also be a high number of households with a head of household under 60 without any children. In earlier versions of BREDEM, the presence of any children triggered an additional amount of predicted non-heating energy use in the home.

However, when investigated further, it is much more likely that a household with a head of household under 60 will live in a period home instead of a modern home. This may be because those aged over 60 in 1996 were probably buying their first home during the years of post-war reconstruction with many pre-war homes yet to be renovated, and therefore were more likely to buy a new home. Table 8.3 treats the dataset as a stratified random sample as per the design of the 1996 EHCS:

Table 8.3: Frequency of elderly households living in pre-war households in the fuel sub-sample of the 1996 EHCS

Table of HAGE296X by prewar								
Age of head of household	**prewar**	**Frequency**	**Weighted Frequency**	**Std Dev of Wgt Freq**	**Percent**	**Std Err of Percent**	**Row Percent**	**Std Err of Row Percent**
under 60 years	**Post-1945**	563.0	5278000	324400	37.64	1.860	53.63	2.240
	Pre-1946	634.0	4562000	259200	32.54	1.690	46.36	2.240
60 or more years	**Post-1945**	370.0	2728000	207300	19.45	1.400	65.27	2.990
	Pre-1946	209.0	1452000	148700	10.35	1.040	34.72	2.990

The multilevel model presented in the previous chapter predicts that the effect of household size on non-heating end-use energy is less in prewar housing that it is in postwar housing. Therefore this interpretation of the reasons behind the effect of building age is inconclusive as prewar housing decreases the influence of household size, but the effect of having an elderly head of household also decreases the influence with this group having a tendency to live in postwar households.

Inserting building age as a dummy variable (pre- or post- war) into the algorithm decreases the influence of household size with a slightly higher intercept. This could be because period buildings are more uniform in building form and function, and therefore variation in energy use might be more limited between households of different sizes. The predictor of household size does not distinguish between the types of rooms or the types of people that make up a household in the same manner that different types of materials and boilers are accepted in heating end-uses of a domestic energy model.

One could look at the 2001 ONS Area Classification groups to gain a sense of what types of area have predominantly pre-war or post-war housing. There are some types of areas with a clear

characteristic of having predominantly period housing (two-thirds or 66%): Countryside Communities, Educational Centres, Multicultural Urban, Urban Terracing, and Young City Professionals. These types of areas have in common the characteristics of having more young singles and couples, and a location just outside of the city, town, or village core. Other types of area are typified by the predominance of modern housing: Blue Collar Urban Families, Mature Urban Households, Rural Economies, Urban Commuter, and Young Urban Families. These types of areas have in common a predominance of families and the elderly, and a location further away from the city, town, or village core. The former group also typically have household sizes that are smaller than the latter, but this variable is controlled in the multilevel model.

Table 8.4: Frequency of pre-war housing by ONS area classification group in fuel sub-sample of 1996 EHCS

Table of GROUP_NAME by prewar								
GROUP_NAME	prewar	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
Affluent Urban Commuter	Post-1945	50.00	717300	127300	5.118	0.8885	65.91	5.780
	Pre-1946	44.00	371000	67950	2.647	0.4858	34.09	5.780
Blue Collar Urban Families	Post-1945	137.0	609900	73250	4.352	0.534	72.49	4.040
	Pre-1946	64.00	231500	37000	1.651	0.2695	27.51	4.040
Countryside Communities	Post-1945	3.000	11270	7701	0.080	0.055	26.75	16.65
	Pre-1946	4.000	30850	15550	0.220	0.1111	73.25	16.65
Educational Centres	Post-1945	7.000	58010	29960	0.414	0.2136	29.85	13.34
	Pre-1946	18.00	136300	50650	0.973	0.3604	70.15	13.34
Farming and Forestry	Post-1945	3.000	51910	31490	0.370	0.2244	53.60	17.91
	Pre-1946	8.000	44930	17360	0.321	0.1243	46.40	17.91
Mature City Professionals	Post-1945	18.00	177200	57870	1.264	0.4114	23.34	6.580
	Pre-1946	67.00	581900	96210	4.152	0.6817	76.66	6.580
Mature Urban Households	Post-1945	89.00	648400	118100	4.626	0.8277	71.60	5.550
	Pre-1946	38.00	257200	51530	1.835	0.3697	28.40	5.550
Multicultural Inner City	Post-1945	42.00	233000	46600	1.662	0.3349	49.85	6.780
	Pre-1946	47.00	234400	42280	1.673	0.3055	50.16	6.780
Multicultural Suburbia	Post-1945	39.00	341500	87830	2.437	0.6204	61.48	8.950
	Pre-1946	43.00	214000	58760	1.527	0.4182	38.52	8.950
Multicultural Urban	Post-1945	17.00	101900	42580	0.727	0.3033	22.29	7.920

Table of GROUP_NAME by prewar								
GROUP_NAME	prewar	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
	Pre-1946	49.00	355300	64690	2.535	0.4631	77.71	7.920
Resorts and Retirement	Post-1945	30.00	215900	64200	1.540	0.456	33.94	7.840
	Pre-1946	52.00	420200	75970	2.998	0.5415	66.06	7.840
Rural Economies	Post-1945	36.00	376800	87390	2.688	0.6178	60.96	10.27
	Pre-1946	28.00	241300	87410	1.722	0.6175	39.04	10.27
Small Town Communities	Post-1945	78.00	607100	126300	4.332	0.8817	59.12	6.890
	Pre-1946	53.00	419800	80560	2.995	0.5724	40.88	6.890
Struggling Urban Families	Post-1945	117.0	560800	71990	4.001	0.5225	65.35	4.760
	Pre-1946	71.00	297300	48440	2.121	0.3507	34.65	4.760
Suburbia	Post-1945	43.00	322700	75530	2.302	0.5359	53.14	7.980
	Pre-1946	32.00	284500	61620	2.030	0.4397	46.86	7.980
Urban Commuter	Post-1945	75.00	1307000	197100	9.328	1.3299	87.41	3.000
	Pre-1946	26.00	188400	42390	1.344	0.3041	12.59	3.000
Urban Terracing	Post-1945	20.00	109100	29490	0.779	0.2115	14.95	3.900
	Pre-1946	79.00	621000	87520	4.431	0.6265	85.05	3.900
Well off Mature Households	Post-1945	72.00	970500	153700	6.924	1.0608	64.38	5.540
	Pre-1946	68.00	536900	96280	3.831	0.6809	35.62	5.540
Young City Professionals	Post-1945	12.00	73210	29300	0.522	0.2092	15.32	7.180
	Pre-1946	37.00	404600	153800	2.887	1.0727	84.68	7.180
Young Urban Families	Post-1945	44.00	509400	98670	3.634	0.6962	78.29	6.390

Table of GROUP_NAME by prewar								
GROUP_NAME	prewar	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
	Pre-1946	15.00	141300	45250	1.008	0.3226	21.71	6.390
Frequency Missing = 1								

These are only small clues to why building age as a predictor results in a significant multilevel model based on a predictor of household size and an outcome of non-heating end-use energy. With more accurate location data, one could determine if this effect is simply the similarity of the location of homes. With more studies on the energy use of people of different ages, one would begin to understand more about why post-war housing with a higher predominance of families might increase their use of electricity for electronics and appliances faster as household size increases.

Perhaps there are additional variations in larger households a these two groups are different which are having a disproportionate effect on the two different slopes for pre-war and post-war housing in the multilevel model. Before 2005, a different calculation method for large buildings, defined as at least 450 square metres, was in place because of limitation in the predictive power of SAP2001 (BRE, 2001). This requirement was removed in 2005, but this was due to improvements in the modelling of the two heating zones as the algorithm for non-heating end-uses did not change.

The effect of building age on non-heating end-use energy is a difficult predictor to unravel. At the outset of this research, building age was not hypothesised to have an effect on this kind of energy use when controlling for household size. The first proposed explanation of a connection between the age of the occupants and the age of the buildings was inconclusive. After examining the profiles of the 2001 ONS area classification groups, there was a possible connection between the tendency to live in a modern house and living as a family unit with children involved with some additional effect on non-heating energy use, but there also location factors to consider as post-war housing tended to cluster on the edges of urban areas. One historical narrative describes a push in modern (prewar) architecture to suppress the domestic, encouraging people to a world of adventure and change, but contemporary (postwar) housing as a reaction to this, espousing the

"pleasures of stasis" in both architectural and urban design (Harr and Reed, 1996). Conversely, pre-war housing clusters together immediately outside of the city, town, or village core around public transport routes. Without either exact locations made available to researchers or focused surveys of metered energy use of neighbourhoods that represent relevant area typologies, these issues will continue to challenge researchers working on energy use of non-heating end-uses in households.

8.3.4 Conclusions

The alternative solution of the multilevel model as a base for a bottom-up domestic stock model of non-heating end-use energy is one that needs further exploration. The model depended on the use of the interaction term to interpret fixed effects coefficients as absolute increases in energy use instead of representing a relative increase. The need to maintain this interpretation also meant that area typologies that were represented by area classification groups were not differentiated in the model by the range of 40 characteristics that created the rationale for the group, but instead by the mean interaction term for that group. Other potential predictors were limited to binomial "dummy variables" that in practical terms created two models, one for each of the binomial cases that were still based on absolute differences in energy use. The one "dummy variable" predictor tested that proved to be both significant and measurable (related to the dwelling and not to immeasurable occupants) was building age, which triggered an investigation into the reasons for building age being a predictor of non-heating end-use energy as opposed to heating end-use energy. This investigation came to the conclusion that there were possible family structure and location factors for which building age could be standing as a proxy, but without the release of locations of households surveyed in national fuel samples of housing surveys or targeted surveying of energy use of households in representative neighbourhoods, these factors are too difficult to estimate.

8.4 Further possible models of non-heating end-use energy

8.4.1 Introduction

There were approaches that were considered but not taken forward in this thesis, as well as further approaches that could be possible by taking on further specialist analysis skills from sectors outside the built environment sector. The first approach is to use what is called hierarchical related regressions to use the mean value of the interaction term as a predictor, but the values of characteristics that make up each area typology. The second is to run the single-

level model using robust regression, a non-parametric technique to estimate energy use. The third is the archetypal method of profiling households as a reduced method that assumes that housing surveys will not improve, to provide more detailed information. The fourth method is to attempt to return to a full version of conditional demand analysis for electricity-using devices in a household.

8.4.2 Hierarchical related regression

Hierarchical relation regression is a form of multilevel modelling that claims to allow researchers to model individual and aggregate data simultaneously, while including information on the dependent variable at the aggregate level (such as the mean electricity use of an area classification group) as well as data from aggregation units not available at the individual level (such as the housing market or the range of non-domestic uses in the area). This method would not likely succeed without more detailed location information, with the LLSOA or similar census area to which a household belongs attached to the survey data.

Investigation of how to combine individual-level and aggregate-level data has been continuing ever since the ecological fallacy was identified by Robinson in 1950 and emphasised for the built environment and geography community in 1984 by Openshaw. Two research projects have taken place in the last decade – the first was by a political science group focused on reconstructing individual behaviour from aggregate data at Harvard University (King et al., 2004). A second biostatistics group which will be discussed here is the BIAS (Bayesian methods for combining multiple Individual and Aggregate data Sources in observational studies) project at Imperial College London (Jackson et al., 2008). The graphic below is a visualisation of the Imperial team's proposal for hierarchical related regression as an alternative to multilevel modelling:

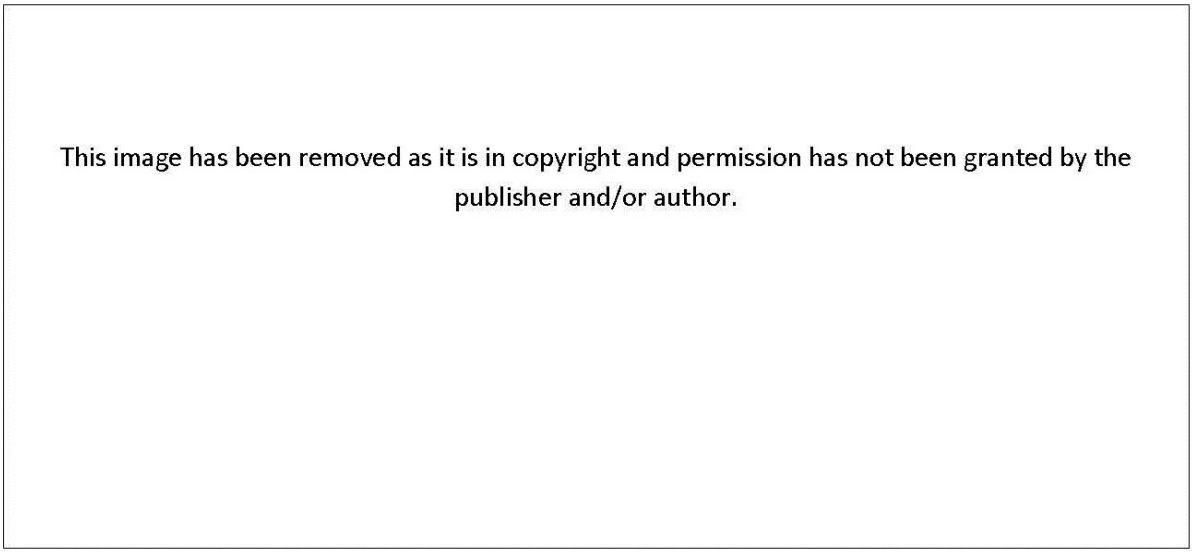


Figure 8.4: Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors (Jackson et al., 2008)

Given data is denoted as squares and unknown quantities as circles. The aggregate (y_i) and individual (y_{ij}) outcomes in this type of model are assumed to be generated by the same underlying individual-level model, in terms of predictors $\mathbf{x}_i$ (which the team interprets as area level effects) and $\mathbf{z}_i$ and $\mathbf{z}_{ij}$ (which are interpreted as individual level effects). The underlying individual-level model has a number of fixed effects at the individual level with intercept μ and slopes $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, with a fixed effect at the group-level γ. The top right-hand side of the illustration proposes that the within-area probabilities φ_{ik} of predictors that are measured at both the individual and aggregate levels are estimated by using the mean value of the predictor at the aggregate level $\mathbf{z}_i$ and samples of individual level predictor $\mathbf{z}_{ij}$ (as an example, the Samples of Anonymised Records made available to researchers by the Office for National Statistics). The bottom part of the graph illustrates that the unknown outcomes from the individual survey data in years other than the target year are imputed by predicting from a regression model fitted to the values in the equivalent survey data at the individual level for the year of investigation.

This method used maximum likelihood estimation as the product of the likelihoods for the individual and aggregate components of the model illustrated. To combine the individual and aggregate data in this fashion, the team states that one must assume that the "exposure" – in their case, exposure to a medical treatment – to outcome relationship is the same in the sample of individuals as in all of the individuals comprising the aggregate data. "Exposures" in the

contexts of non-heating end-use energy could be local authority policies, efficiency awareness campaigns, or correction factors for low-energy lighting fixtures or daylighting that are already in SAP for new-build housing. Maximum likelihood estimation is also used in the multilevel model that was used in this thesis, so this type of extension to the work could be quite valuable. However, there is a downside that if predictors are used that are not measured at both levels, the results will only deliver predicted relative, or contextual, effects of the "exposure" instead of predicting absolute differences in energy consumption between households.

8.4.3 Treating the data as non-parametric: Robust regression

Another technique that was not employed but may be possible with additional specialist knowledge is to use a family of modelling techniques that are collectively called robust regression. These techniques are designed to lessen the sensitivity models have to outliers compared to classical linear regression. They are also iterative in nature, meaning that they require modern computing power both to run and interpret. Robust regression would retain households that do not use electricity for heating, causing Type II errors through the inclusion of high electricity users connected to heating and not household size. If there were more data points available, there would have been less need for seeking out non-parametric methods.

In the single-level model, the response to having outliers, defined as more than 2 standard deviations from the mean of the transformed dependent variable of non-heating end-use energy, is to remove them under the assumption that they could not be cases that did not use electricity for heating. As this was a conservative classification with a large (more than 3,000 case) dataset, a good argument could be made for a more lenient classification and therefore less of a need for non-parametric techniques.

One example of a robust regression technique is M-estimation (Huber, 1964) which is implemented in the SAS software package as PROC ROBUSTREG (SAS Institute, 2011). M-estimation uses a technique called IRLS (Iteratively Reweighted Least Squares) which is carried out inside an iteration loop. For each iteration, a set of weights for the observations is used in the least squares fit. The weights are constructed by applying a weight function to the current residuals. Initial weights are based on residuals from an initial fit of least-squares estimation using classical linear regression of each case, and the regression is repeated, which gives rise to new weights, until the regression coefficients converge (Fox, 2002).

This technique would protect the model from undue influence from unusual data when there is no reason to suspect that the data is invalid. The reasons cited for not choosing robust regression are that the technique is computationally intensive, and that the standard errors become very large, which reduces the confidence in parameter estimates compared to classical ordinary least squares linear regression if there is data that can be transformed to be approximately normally distributed and pass the parametric tests (Field and Miles, 2010b).

8.4.4 Archetypal method

The archetypal method was not used because it was not a practical solution to the problem of estimating non-heating end-use energy in an existing home. Archetypal methods, with the advance of technologies such as LIDAR (Light Detection And Ranging), satellite imagery, and automatic recognition software, could come back into the range of options available to the researcher. At present, the requirement of the archetypal method to estimate the layout and configuration of rooms could not be used as predictors in a regression analysis, as the archetypes were akin to categorical, and not ratio variables.

The archetypal method was also deemed to be questionable for use for the estimation of non-heating end-uses in the residential sector because the justification for creating archetypes is for estimating indoor air temperatures and the resulting heat demand (Parekh, 2005). Previously some reviews of domestic energy stock modelling in the context of Europe have been presented, and all of them are centred around the heat demand of the dwelling with non-heating end-uses only occasionally contributing to the model by their heat emitted from metabolic processes or the operation of electrical and cooking equipment (Kavgic et al., 2010).

New detailed surveys of the use of a building made for the commercial market might provide relevant detail for either the archetypes or for the missing aggregate building age data for verification of a multilevel model. These surveys may also estimate the residential and non-domestic components of complex buildings and areas. Recently, there have been major advances in the amount of three-dimensional mapping available, notably with the emergence of the Cities Revealed project and more recently the UKMap projects (Groom, 2009). These might be able to not only describe the size of a dwelling, but more accurately describe the concentration of different dwelling types or even estimate building age from survey data collected on the ground for mainly urban areas throughout the UK. However, the pairing of these types of surveys with addresses covered by the fuel sample of the 2011 English Housing Survey is unlikely to occur, and even less likely with addresses with meter data already held by energy companies.

8.4.5 Return to conditional demand analysis?

The lack of reliable predictors for domestic non-heating end-use energy may lead to a call by some in the research community to return to conditional demand analysis (CDA) techniques discussed in the literature review section of this thesis. This was used extensively by building researchers in the 1980s in the modelling of households themselves using survey data for a small number of households (Parti and Parti, 1980, Uglow, 1982) and was the previous validation model for the housing stock model for non-heating in BREHOMES from the 1990s (Environmental Change Institute, 1995). When CDA was first proposed by energy researchers, the list of possible appliances was tiny compared to today. In addition, there were assumptions that could be made about the link between ownership of a certain "basket" of appliances and these people's socioeconomic situation. With the progressive reduction of cost of appliances and electronic devices relative to earnings, the ownership of these goods may now be one that can only be modelled by culture and choice rather than by affordability.

The typical "basket" of goods that a household owns has grown substantially, with a number of new devices to keep track of bringing the total number to a level and a pace of change that must surely be unworkable for a researcher to model using CDA. Some researchers have turned to neural networks as an alternative to measure human tendencies to use and acquire devices (Aydinalp-Koksal and Ugursal, 2008). A major study into the energy consumption of households in the United States revealed a pattern that is surely the case in England and throughout the developed world: while most "white goods" or major appliances have become more energy efficient, the average household now has many more consumer electronics than before. The study summarises the situation of the rise from 1978 to 2009 of personal computers, televisions, and a loose group called rechargeable electronic devices composed of items such as mobiles and music players. In short, personal computers went from a standing start to 76 percent of homes, one television went to 2.5 televisions per household, and one-third of all households in 2009 had four or more rechargeable devices (U.S. Energy Information Administration, 2011).

The other sector of conditional demand analysis that has not been fully explored is the link between the active occupation of a dwelling and its electricity use. Some households may be in circumstances that allow them to spend relatively larger amounts of the year away from home on foreign holidays. Others may choose to spend more of their leisure time under normal conditions away from the dwelling by, for example, eating in restaurants or taking in more movies and concerts. Others may choose to take on additional employment – the UK has the longest working

hours in Europe. Per hour of occupation, these households may be spending energy at a rate far greater than their peers, but because models such as the one explored in this thesis measure electricity on an annual basis (following the frequency of available data), this rate of usage is masked.

At the other end of the spectrum, there are those who spend large amounts of time at home compared to others such as the underemployed, the elderly, students, or families with young children. They may consume more energy on an annual basis, but measured at an hourly rate, these groups could well use it more prudently. If they went to work or on holiday instead of caring for family members or "staycations" would the energy use in those buildings or processes personally assigned to them actually be less that the extra energy spent in a domestic setting? These are questions about the nature of contemporary society and its relation to energy consumption both inside and outside the home that, if CDA is to return, should be addressed in a much more serious fashion than previously.

The survey work that would be necessary to bring about a return of conditional demand analysis would have to provide a powerful enough sample of the population and be extended to cover the actions of individuals and not just households. This would involve the combination of smart meters and smart plugs in households at shorter time intervals than the traditional half-hour, with time use diaries that are also detailed enough to be able to differentiate the actions of different individuals in the home. If location factors such as proximity to leisure activities or long commutes to work are to be considered, the exact location of the participants would be needed. These types of surveys would be considered to be very invasive, and universities and governments would be hesitant to approve them.

8.5 Domestic energy modelling as a pathfinder for the quantification of sustainable living

8.5.1 Metrics of sustainability

The measurement of sustainable living practices goes beyond the carbon emissions predicted by the residential sector. Defining sustainability in terms of total energy consumption, and even the performance of a building against expectations, is relatively easy compared to defining it as a confluence of social, economic, and environmental factors (World Commission on Environment

and Development, 1987), of which energy use is just one, albeit one that crosses all three of these perspectives.

From an environmental perspective, the study of non-heating end-use energy in households is an instructive study in how to quantify the impact of a household unit on its environment, but there are many other environmental issues such as water and waste as well as social and economic issues. There have been rating scales such as LEED, BREEAM, SB-Tool, and Pearl that have been created to attempt to balance the priorities of each "flavour" of environmental sustainability (Sedlacek and Maier, 2010). They set their own terms of engagement with each category, with a set amount of credits, points, or other types of ranking made available and negotiated against one another in a manner that is more akin to the setting of political or planning policy than a scientific balancing of priorities, as there is no single metric that can be referenced during this process. One can argue that this metric is always the emission of greenhouse gas emissions, or other measurements such as the ecological footprint. The ecological footprint is derived as an annual figure for a defined population by estimating the area of productive land and sea required to support its resource consumption using available and established technologies—examples given are for food, travel and energy use (James and Desai, 2003, Humphrey et al., 2008).

Domestic energy use as an outcome variable to measure the environmental performance of the built environment is equally uncontroversial, but it requires time and investment. It has been far easier to calculate the intent to reduce this amount of energy use through the counting of installations of technologies in the domestic sector. More surveys are needed of actual energy usage and ways to connect them to everyday usage of energy, specifically "elective" energy use unconnected with food or shelter and therefore most of non-heating end-use energy. At present, modelling of non-heating energy occurs between households, and in the case of this thesis, household size.

Other quantifiable outcome variables that offer a measure of environmental sustainability of the residential sector include transport energy use alongside air pollution, water usage and waste generation. These are all important to the environment and important areas for investigation, but agreeing a common metric that can compare the cost to the environment of all of these variables is difficult and has, it seems to me, a political as well as a scientific decision. Currently, the importance of each of these variables alongside domestic energy use in comparison with each other is vague, and conflict between the policies recommended by research in each of these

areas goes unresolved. The agreement of a common metric might lead to perverse consequences, It may tip the balance towards or away from some of these areas from the current status quo, causing reallocations of research and development budgets and upheaval throughout the sustainability research complex. This makes it unlikely that such a metric will be agreed, but informally the lay public recognise carbon dioxide emissions as this metric.

This metric of absolute carbon emissions is, however, misconceived when thinking about the types of effort, money, and inconvenience tolerated to enable reductions to happen. This introduces the issue of the quality of the reduction in energy use, water, waste, or any of the other metrics of environmental sustainability as opposed to the quantity of these reductions. This quality is the amount of effort and cost that it may take to achieve the same quantity. Some studies have been carried out previously to identify the payback of different types of technologies for the reduction of heat energy demand within the context of England (Ritchie and Thomas, 2009).

In domestic non-heating energy use, there has been a reduction of energy demand by white goods such as refrigerators, stoves, and washing machines after the implementation of the European directive on labelling of the energy consumption of appliances in 1992, which was revised in 2004 (CECED, 2011). Once the labelling system (A-G) was in place, it enabled consumers to quickly determine the energy efficiency of these large appliances, and the average energy efficiency of appliances on sale in the European Union rose by 15% within 15 months (Parliamentary Office of Science and Technology, 2005). This shows the power of clear information on the amount of electricity use in the control of the consumer. However, this kind of labelling does not extend to electronics, computers, or rechargeable devices, and these are the new drivers of non-heating end-use energy as demand increases by around 2 percent per year. Furthermore, increases in energy efficiency results in a rebound effect (Wilhite and Norgard, 2004).

8.5.2 Transfer of knowledge of behaviours of non-heating energy use to understand attitudes to sustainable lifestyles

The measurement of non-heating end-use energy is crucial for the understanding of sustainable living because of our increasing dependence on technology to deliver information and services within the home, which in turn requires increased energy use. In time, the fixed effects of energy use that are associated with basic human survival (food and shelter) will become less significant as heating and appliance energy efficiencies are increased, whilst the more variable and fast-

changing effects associated with electronics and gadgets will become more influential (International Energy Agency, 2009). These behaviours are also taking place in the most private part of people's lives, which is the home. Other behaviours relating to sustainable lifestyles may be greatly affected by the need to project a positive self-image to others and therefore, in the words of the charity Global Cool, "Doing it in public" as opposed to following through with people's stated behaviours in private (Global Cool, 2011). Non-heating end-use energy is a powerful metric for understanding the differences between public posture and private action, or stated and revealed beliefs that relate to sustainable living practices. The more that these behaviours are understood and the researchers are better able to predict future energy demand in households, the more that researchers in other disciplines will be able to understand the extent to which people are living a more sustainable lifestyle.

The amount of energy used for these end-uses that are largely unconnected with survival is a reflection of one's willingness to use resources, one's access to them, and their affordability. When a person uses more resources, this may be in pursuit of various personal goals related to work, leisure, and status within various social circles of colleagues, peers, and family. In a bottom-up domestic stock model that takes the household as the primary unit – essentially, a single household energy performance model – these behavioural factors can be explicitly built into the model. As a unit intended to deliver the expected consumption of the residential sector, any details about the variability of usage within the household become blurred by assumptions about the average or typical behaviours of - in the case of this thesis - households of a certain size. In addition, the physical elements of domestic energy modelling – correction factors for lighting in this case – have been measured extremely carefully, and it is hard to ignore the hard work of a generation of experts in building science in favour of something that is difficult to define and more difficult to measure, record, and predict for the future.

If these behaviours within the home that relate to these "elective" energy usages without the gaze or judgement of their peers or colleagues, can be accurately predicted, then perhaps other behaviours that are damaging to the environment relating to procurement, food, leisure, and transport may be better understood. The results can also be a better diagnostic of revealed, instead of stated, beliefs about sustainable lifestyles than energy for heating end-uses. It has been found in previous studies that stated pro-environmental behaviours are not a good predictor of energy efficiency (Cornelissen et al., 2006). The use of domestic non-heating end-use energy is also not connected with social norms; however, the possession of consumer electronics

(despite gains that may be made in the energy rating of the individual device) are a growing social norm in industrialised countries. This conflict between public showing of status and private consequences of environmental harm through increased electricity use, if properly investigated, would be enlightening for us in other areas.

The way that this thesis begins to approach this is by the use of the neighbourhood unit as one that can start to explain these internal variations. There has been evidence in the past that social capital, or the sharing of experiences between people in the same community, can penetrate beyond the walls of the home (McMichael, 2009). The area classification system accentuates the similarities between households within the same classification as well as the dissimilarities between households that belong to different classifications. It is a crude tool that swamps individual level meter readings with mean energy use data at the group level instead of a detailed study of the variation between households and their characteristics. It does, however, start to predict variation in households of the same size based on the type of group to which they belong.

These classifications give the researcher some clues as to the effect of local social and economic conditions on non-heating end-use energy while allowing for variability within the group. However, they can only describe characteristics associated with groups that have large mean household sizes within the context of this research. Other research into sustainable lifestyles may use a person as the unit of measurement using different independent variables such as income or age to predict an outcome. This may add a third level into the multilevel model, so there will be variation accounted for between people within the household, between households within a group, and between groups. Another dimension of research may be "within people" or the measurement of the scale and rate of take-up of sustainable lifestyles over time.

8.5.3 Quantification of sustainable living as part of general assessment of the planning of residential communities

Sustainable living can often lead to a simple quantitative analysis of resource consumption. When these goals come into conflict with more qualitative goals of social inclusion and design, then the decision-making process of what to build and where to build it becomes negotiated and contested through urban planning mechanisms and the political process (Guy and Marvin, 1999). In England, there are set criteria for the design of communities that should be adhered to by anyone proposing a development backed up by a design and access statement (DETR, 2000). These criteria are balanced against each other in a quasi-legal fashion. A quantitative measurement that balanced all of these factors using local and national priorities to aid decision-

makers in the future should be considered as a transparent measure of a proposal's own performance against the criteria set down for it and attempting to balance priorities, raw performance, and ease of achieving this performance for any development.

The input of energy use for England is one of many and linked to a tiered standards for the predicted dwelling emissions rate called the Code for Sustainable Homes (Department for Communities and Local Government, 2008a). Like the European Directive that created a six-point rating scale for appliances, the Code for Sustainable Homes is also a six-point scale. Achieving a certain level in the Code requires points earned from reducing resource demand from energy, waste, and water. The Code is a single, national code that applies equally to all housing in the future and is the basis for compliance with the stated goal of all new buildings to be zero-carbon from 2016 in England (Department for Communities and Local Government, 2007a) and the European Performance of Buildings Directive 2010 which requires all new buildings to be "near-zero energy" from the year 2020 (European Commission, 2010).

Standards for energy reduction in the residential sector are uniform and equally applicable to all housing. However, the ease of achieving near-zero energy may be different from location to location depending on both orientation, local energy supply chains, occupancy, and usage patterns within the home. Likewise, other more "soft" standards, such as ease of movement, quality of the public realm, or the adaptability of the building for alternative land uses (DETR, 2000) present varying difficulties depending on local conditions, including if high-performing housing is a norm within that neighbourhood.

Even though the weighting of different measures of sustainability and quality against each other would be an intensive political and technical process, it would be instructive for the planning system, stock modelling, and the understanding of the general public. The general public currently is bombarded with statements about a wide variety of energy-saving installations or measures as well as other measures to improve the quality of the built environment. The total benefit is masked if they are measured by a simple count of measures, which encourages measures that are small and numerous rather than large and time-consuming. In addition, the benefits could be included in an effective argument that emphasises the points that are personal, tangible, and immediate if they are done for quality reasons in addition to mitigating over-use of energy and subsequent climate change.

8.6 Summary of Discussion

This chapter discussed the ramifications of the research on the domestic stock modelling of non-heating end-use energy in the context of England. First, the single-level model proposed in previous chapter offers a simple, verified solution using the same variables and housing survey dataset to predicting this type of domestic energy use, yet some issues around the extremes and collecting of data remain. The multilevel model provided an alternative approach by allowing variation between groups, but the causation was limited to mean values of the same variable at the individual level instead of using other variables to measure between-group variation. Other models of non-heating end-use energy were argued to have potential in the future, from ecological regression to a return to sub-metering individual items of electronic equipment for data. Finally, this type of energy use was discussed as a possible microcosm for the understanding of human motivations, both stated and revealed, for sustainable living as a confluence of social, economic, and environmental factors. For these reasons, further research and gathering of data is needed to measure the energy requirements of households for these end-uses.

Chapter 9 - Conclusions

9.1 Introduction

Creating a bottom-up domestic energy stock model of non-heating end-uses is a difficult task and very different to creating a stock model of internal domestic heat demand. It is a task that does not seek to predict energy demand based on building physics, but on an analysis of the combination of energy demand data and the physical and occupant dimensions of a household. The interaction of the built and social environment of households and, in this work, neighbourhoods, is the defining characteristic of domestic energy modelling of non-heating end-use energy. Although a single-level model of energy use using this term is still relevant and can be used immediately, there is significant long-term potential in using a multi-level model of this kind of energy use using households at the individual level and neighbourhoods at the group level. This type of modelling also has the potential to show the way towards the quantification of the environmental characteristics of neighbourhoods and lifestyles, if data collection is better targeted and statistical techniques applied to the built environment sector becomes more sophisticated in the future.

9.2 The real potential of the multi-level model

A major contribution to knowledge from this work was a domestic energy stock model of the residential sector of England for non-heating end-use energy that measured a "structural deficit" of energy use associated with neighbourhood membership. The type of area that contained a household explained almost half of the variation of non-heating end-use energy using household size as a predictor. After further examination, the model was relevant to nearly all types of areas, with two significant and interesting exceptions. However, the current aggregate census data available is not quite adequate to verify against aggregate energy use statistics, and further targeted measurements of multiple homes in a neighbourhood should be carried out in the future. These three conclusions are considered in turn.

9.2.1 Domestic energy stock modelling using group and individual level predictors

The creation of this multilevel model changes the way that domestic energy stock models in England take into account drivers of non-heating energy use in the residential sector. The change in the model now considers the area context of households as an equally important explainer of variance between households as the household itself. This has some potentially profound

modelling and regulatory implications if the neighbourhood of the dwelling is considered as a predictor of emissions as well as the dwelling itself. This could cause considerable public debate on the responsibility for future carbon emissions of occupants, constructors, architects, and engineers in the development of housing.

A multilevel approach uses both the individual household size as defined by the interaction term composed of the number of occupants and the number of rooms and the mean household size of an area classification defined by the Office for National Statistics for Lower Layer Super Output Areas in England as predictor variables for the outcome of annual household non-heating end-use energy. This has the advantage of having real dimensions, measured in kilowatt-hours, of group-level and individual-level effect sizes. It has the disadvantage of being limited to the same groupings of independent variables at both the group and individual level as there are other predictor variables that the researcher should wish to explore that are only available at one of the two levels. This was partially demonstrated by the inclusion of individual-level building age as a binomial predictor to explain more the variation of individual household non-heating end-use energy by making significant the interaction between mean household size at the group level and household size at the individual level.

In order to test the viability of area classification and mean household size as an effective group-level predictor of energy use, more targeted surveys of electricity use in buildings without electric heating need to be commissioned by either the academic community or central government. The weakness of the current framework is that without location data, characteristics unique to each neighbourhood will be masked when agglomerated into area classifications. The area classifications as well are only proxies for occupant behaviour and generalise the socioeconomic influences on energy use from the very different lives that individuals lead as occupants of their individual households.

These surveys could reveal much about individual and shared patterns of occupant behaviour in relation to energy use in buildings if a representative sample of housing within each neighbourhood that is represented by a Lower Layer Super Output Area in England. Area-based demand reduction campaigns (Energy Saving Trust, 2009a) could be better refined without the need to buy in market research and credit data from private agencies where the methodologies and raw data are not in the public domain or undergo scrutiny from the research community. The individual neighbourhood itself could set appropriate local energy demand reduction policies and microgeneration policies based on the information collected.

9.2.2 Characteristics and trends of energy use of different area classifications

There are characteristics, instead of predictors, that can describe trends of increasing mean household size and energy use of different area classifications. The exceptions to these trends merit some attention for those who are looking at possible causation between the external built environment and occupant choices of non-heating end-use energy. As housing density increases, then household size correspondingly decreases, except for areas labelled as “multicultural urban” which are high density and high household sizes, “small town communities” and “resorts and retirement” classifications which have low density and low household sizes. The same groupings hold true for public transport ridership levels which are mostly closely related with urban density. A second characteristic is that as mean household size decreases, the proportion of flats increases, except for, again, “multicultural urban” which are medium-level flat proportion and high household sizes. A third characteristic is as household size increases, the amount of socially-rented housing decreases, except for “multicultural urban” and “countryside communities” classifications which have a medium-level of socially-rented housing and high household sizes, and the “resorts and retirement” and “educational centres” classifications which have a medium-level of socially-rented housing and low household sizes.

These area characteristics are those that could describe group-level predictors of non-heating end-use energy in future domestic stock models. For architects and planners, this is an exciting and challenging approach to the prediction of energy use at the neighbourhood, city, regional, and national scales. These are prediction factors that are included in guidelines for building sustainable cities for reasons other than energy use in the home – such as reduced energy demand for transport use from increased public transport usage, reduced energy demand for heating in blocks of flats due to the stack effect, or increased energy infrastructure efficiencies from reduced distances between dwellings. In the future, one could apply these principles to the prediction of energy use internally to the home changes how urban planning connects sustainable urbanism and green buildings beyond those already quoted in design literature on the connection between daylighting and use of artificial lights (Homes and Communities Agency, 2007, Department for Communities and Local Government, 2009, BRE, 2010).

These “exceptions” deserve extra attention in the setting of strata in housing surveys (Department of the Environment Transport and the Regions, 2002) as well as targeted studies of individual neighbourhoods of these area types. For the urban planner, this should especially be true of household in the “multicultural urban” type which has been the subject of first that of a desirable

urban lifestyle (Jacobs, 1962, Hall and Ward, 1998) followed by claims that diversity of urban form and uses are desirable and sustainable in the writing of British urban policy (Department of the Environment Transport and the Regions, 2000a, Department for Communities and Local Government, 2009).

9.2.3 Building and neighbourhood age as aggregate statistics

In the multilevel model, adding building ages as a binomial variable created a statistically significant interaction between the mean household size at the group level and the household size at the individual level in the form of two separate algorithms for pre- and post-war housing. As dwellings get larger, energy use grows in pre-war housing at approximately 20 percent of the rate of post-war housing. This was a moderately surprising result. Period and contemporary housing do exist in different configurations, architectural styles, wiring systems, age of householders, and urban locations. These both reinforce the socioeconomic and built environment characteristics that are the basis for area classification and the trends of dwelling choice from different types of socioeconomic groups and family sizes.

It is certain that information for all individual dwellings in England by the Valuation Office Agency (VOA). If this data can be made available, only at the aggregate scale and only at the categorical or even using the binomial variable of pre-war and post-war used in this thesis, verification of the multilevel model would be possible and further advance the understanding of non-heating end-use energy in the residential sector for domestic energy stock modellers using this method. In England, the data is beginning to become dated as revaluation last occurred 20 years ago, even with a programme of readiness should any government wish to trigger a new one. The reticence to do this by successive governments shows that they are not willing to release the data that determine the rate to council tax payers, let along researchers or other interested parties.

There are some urban areas that have attached building age data to individual addresses from desk and visual surveys available from the academic community (Landmap Service, 2011). The information at the aggregate scale is collected in different ways. There is a privately-funded database called National Building Class that is currently being built that contains building age generated from aerial scanning and photography (Geoinformaton Group, 2012). HEED may also be a future source of aggregate information as it collects information from upwards of 10 million dwellings. These have the potential to fill some of this gap in the data over the next decade if the VOA data is not made available. In addition, it would be useful to compare the model using an individual level variable with one at the group level – in essence, an estimate of neighbourhood

age. With complete aggregate building age data, the calculation of overall neighbourhood age could be a valuable group-level variable to use in place of aggregate individual building age.

9.3 The single-level model is still relevant

A second contribution to knowledge was the revision of the single-level model of non-heating end-use energy. The single-level model of domestic energy consumption, with the interaction term composed of two independent variables – the number of occupants and the physical size of the household – predicting energy use of a household, is still relevant and applicable to domestic energy stock modelling. This model is still solid and can be applied to England immediately. This entails a change in variable transformation from the algorithm currently used in England, a mechanism for an annual update of the model, and aggregated data for both the dependent and independent variables for verification of the model. These factors can guarantee a robust single-level model in the near future before a presumed transition to multilevel methods.

9.3.1 Variable transformation

There were two variable transformations that occurred in the single-level model that differentiates it from that embodied in the Standard Assessment Procedure currently being used in England. The first transformation is the exchange of the number of rooms for usable floorspace in a household as the independent variable representing physical household size that composes part of the interaction term. The second transformation is the square-root transformation of the dependent variable of non-heating end-use energy instead of using a logarithmic transformation. These two moves in the setup of the model more closely align building form and energy demand together, but also enhances, not diminishes, the effect of large household sizes. This thesis concludes that these two approaches are correct, but at the extremes - extremely large households or complex situations such as overcrowded dwellings – the model loses predictive power. With the current amount of data available on these types of cases, desk-based methods will continue to falter unless targeted in large scale surveys. Unfortunately in England, the current methodology for the English Housing Survey has removed stratification and therefore oversampling of extreme cases (Department for Communities and Local Government, 2011b).

Moving from usable floorspace to the number of rooms per household may seem like there is a loss of precision, but statistically they are equally suitable for the model, and more data is available for the number of rooms of all dwellings as opposed to their usable floorspace. Early studies on the number and frequency of lights and appliances in buildings using conditional analysis methods

suggested a connection between the form of a dwelling, including the frequency of new rooms, their use as living spaces, bedrooms, or kitchens. Usable floorspace as a measure of physical dwelling size treats it as a formless block of space, whereas the number of habitable rooms, although undifferentiated by size and use between each other, explains the increase of non-heating end-use energy in a different, and, judging by the coefficient of determination (R^2) equally valid approaches to physical household size.

The square-root transformation of the dependent variable is another major change driven by the resulting transformed dataset, derived from the 1996 English House Condition Survey fuel sub-sample, modified by the 2008 Living Costs and Food Survey, becoming approximately normal. The resulting algorithm is much different in its treatment of large households as opposed to the alternative transformation currently used by the Standard Assessment Procedure in England – households increase their use of non-heating end-use energy at a faster and faster rate as they get extremely large instead of levelling off at a horizontal asymptote that was first proposed in energy modelling in the 1980s. There does need to be more targeted measurement of extremely large households to get a better sense of what drives this energy use and if either conclusion holds true. Mathematically, it would be wise to leave out extremely large households as exceptional and *sui generis*. However, as homes with right or more rooms comprise around ten percent of the housing stock in England (Office of National Statistics, 2005), the amount of energy these households use is potentially very large that the predicted energy demand of the entire residential sector would continue to be significantly below the actual energy used.

9.3.2 Annual updating

This model makes a significant contribution to domestic energy modelling by concluding that the single-level model can be, with some confidence, be updated annually and not require a fresh intensive survey of electricity meters in households to be accurate and relevant. The additional data is self-reported electricity bills in the Living Costs and Food Survey instead of actual electricity use, but with the baseline meter data from the 1996 English House Condition Survey, errors and inaccuracies can be minimised. With the rapid growth of demand from lights and appliances predicted by national and international forecasters (International Energy Agency, 2009, Ekins and Dresner, 2006, Energy Saving Trust, 2007a) the need to update with partial data every year is both needed and, currently, not possible.

9.3.3 Verification

The other advantage of annual updating of this model leads to another significant contribution – verification. There is aggregate data available in England from their central government to verify the predicted non-heating end-use energy of the residential sector of a large and representative portion of England for any particular year since 2006. In addition, the household size in terms of numbers of occupants and numbers of rooms is available in each census area from the 2001 United Kingdom Census which also can be updated with population projects made annually by the Office for National Statistics. For the year 2008, a number of census areas were selected based on their maximisation of central heating in that district and the model over-predicted the non-heating end-use energy of their residential sectors by less than 1%. This kind of verification would have not been possible without first the updating of the model to the year 2008 or the change from usable floorspace to the number of habitable rooms to be able to access the aggregate data available for each census area.

A future requirement to install smart meters in the UK looks unlikely to lead to better verification. A major consultation about access to this data is currently underway (Department of Energy and Climate Change, 2012). At present, there are no plans to match smart meter data to other government surveys such as the English Housing Survey. Aggregation and anonymising of data is still proposed be permitted by the government for gas and electricity networks to collect and disseminate in a similar fashion to the current LLSOA statistics. There are provisions for homes to be automatically enrolled in energy saving-trials unless they opt-out to avoid survey bias, which could, if the rough location and capacity of such schemes were made available, help measure the impact of microgeneration on electricity demand and aid in the understanding of future data collection of aggregate electricity use. However, this might also be deemed to be private information and uncollectable with location data, creating unknown bias in certain areas when collecting aggregate energy use in small areas.

There were variations in the accuracy of the model dependent on the type of area which leads to the conclusion that a multilevel approach should be ultimately adopted in the near future. Rural or well-off suburban areas were underestimated in comparison to actual energy use. Inner urban or less well-off suburban neighbourhoods were overestimated. As this thesis has already indentified large households as an area for further investigation, it could be that large households in well-off areas use so much more energy than large households in poorer neighbourhoods that they need to be considered separately in the model – yet more evidence of a need to move to consider group-

level effects. There are several physical variables at play that need to be investigated further, such as people per room, average size of rooms, appliance density per room, and the energy rating of appliances. There are also social variables that emerge from human to device interaction with different levels of correlation with dwelling occupants. This leads to questions around verification of any model being able to take account of not only the sustainable design of buildings, but also the lives likely to the led within them.

9.4 Domestic stock modelling of non-heating energy as a measure of sustainable neighbourhoods and lifestyles

Measuring the socioeconomic characteristics of different areas does, in small part, offer a small window on why people decide to use energy, whether inside or outside of their home, if it is available to them and, unlike heating, not directly connected with the shelter requirement for basic human survival. Non-heating end-use energy is a clearly measured, continuous variable with a clear conversion to a carbon dioxide equivalent depending on fuel for electricity generation. Other metrics for sustainability do not have this sort of structure of how to measure usage or the impact on the local environment up to a planet-wide ecological system. This has lead to rating systems for housing that juggle different environmental, and also social and economic, priorities for determining what has achieved a “gold standard” for sustainability by re-scaling independent variables and the establishment of thresholds for the outcome variable for the achievement of different levels of sustainability. The future direction of this kind of research into the influence of neighbourhoods on the sustainability of buildings is to challenge the current rating systems and codes with one that has sustainability as a continuous outcome variable. Using energy as a spearhead and the multilevel method as a modelling process, the way sustainable living in the city is measured and “awarded” could be challenged.

9.4.1 Challenge to current rating systems with a categorical outcome variable

The current housing rating systems available throughout the world such as BREEAM, LEED-H, SBTool, or the Code for Sustainable Homes for England, have “points” that can be earned in a series of categories, one of which is energy use. In the case of England, energy use is measured by the Standard Assessment Procedure rating. How much energy, water, waste, etc. is predicted to be reduced by the building to earn a ratings point is up to the methodology of each rating scheme. Each of them will have sustainability priorities set within the national context in which the rating scheme was created.

These rating systems are often applied within radically different local contexts, even within one nation, especially those that are have major urban-rural and climatic divides. However, the method for "earning" rating points and the final points thresholds for code level "4" for Code for Sustainable Homes or a "silver" award for LEED remains the same. Re-scaling the points for each setting or local authority priorities is time-intensive and open to improper alteration by local planning authorities.

9.4.2 Sustainable communities measured by a continuous outcome variable

Following the lead of this work on energy use, the influence at the group-level as well as the individual-level variables of sustainability can improve measurement of human environmental impacts. As an alternative approach, a continuous outcome variable of sustainable living can be predicted from of the interaction of all of the independent variables of sustainability. There are some calculations of global footprint as an amalgam of different metrics of sustainability such as One Planet Living (Bioregional, 2011), but they focus entirely on the choices made within an individual household and judges impact to be equal from the actions of households that are the same size. Instead, an approach that includes local contexts could be piloted in energy use and carbon-dioxide equivalents and then extended to other categories that are involved in the calculation of global footprints or some alternative continuous outcome variable instead of a rating system with less than 10 achievement levels.

The inclusion of group-level factors will also enable governments, researchers, and funders knowledge of variation of individual households from their neighbourhood "norms" and give feedback not just on absolute reductions in energy demand, but also relative reductions to regional, city, and neighbourhood averages and distributions. This is built on the premise that recognition should be based on one's ability to change as well as the amount of "savings" both in energy and in more general measures such as global footprint. This type of standard would be an invaluable contribution to fair and balanced recognition of sustainable living, and multilevel models, with non-heating end-use energy as a pointer in the right direction, should be a standard method in the researcher's skillset.

9.5 Final summary

In summary, this thesis makes several contributions to knowledge about bottom-up domestic stock modelling of non-heating end-use energy within the context of England. First, it evolved the current single-level model of non-heating end-use energy for households as the building block for

domestic energy stock modelling. This made the model able to update itself annually without the need for a fresh intensive monitoring programme as part of a national housing survey. It was a non-linear model that accelerated the predicted growth of energy use in large households as household size gets larger instead of predicting that energy use would reach a maximum. It also used different variables that are replicated as aggregate statistics of the whole population in England to enable the model to be verified against actual energy use. Second, a multilevel approach was developed as the next generation of domestic stock modelling. This methodology found that the effect of area type was almost as powerful an explainer of variance as the individual household when using household size as a predictor of annual non-heating end-use energy. This method, when in a form that can be fully verified using aggregate statistics, can point the way towards assessments of sustainable homes and lifestyles using metrics outside of energy consumption in areas not connected with human survival but of patterns of everyday living.

References

Abercrombie, P. S. & Forshaw, J. H. 1943. *County of London Plan prepared for the London County Council,* London, MacMillan and Co.

Abrahamse, W. & Steg, L. 2009. How do socio-demographic and psychological factors relate to households' direct and indirect energy use and savings? *Journal of Economic Psychology,* 30**,** 711-720.

Adnot, J. 2003. Energy Efficiency and Certification of Central Air Conditioners (EECCAC) Final Report. Brussels: D.G. Transportation-Energy (DGTREN) of the Commission of the E.U.

Alcott, B. 2005. Jevons' paradox. *Ecological Economics,* 54**,** 9-21.

Alexander, G., Warm, P., and Reddish, A. 1983. *Warm and wise: The Energy Matters handbook,* Milton Keynes, UK, Open University Press.

Anderson, B. R. 2002a. *BREDEM-8 : model description, 2001 update,* Garston, Watford, UK, BRE.

Anderson, B. R. 2002b. *BREDEM-12 : model description, 2001 update,* Garston, Watford, UK, BRE.

Anderson, B. R., Chapman, P. F. & Cutland, N. G. 1996. *BREDEM-12 model description,* Garston, Watford, UK, Building Research Establishment.

Anderson, B. R., Clark, A. J. & Baldwin, R. 1985a. *BREDEM : the BRE Domestic Energy Model,* Garston, Watford, UK, Building Research Establishment.

Anderson, B. R., Clark, A. J., Baldwin, R. & Milbank, N. O. 1985b. *BREDEM-BRE Domestic Energy Model : background, philosophy and description,* Garston, Watford, UK, Building Research Establishment.

Angrist, J. D. & Pischke, J. r.-S. 2009. *Mostly harmless econometrics : an empiricist's companion,* Princeton, N.J. ; Oxford, Princeton University Press.

Association for the Conservation of Energy, Impetus Consulting Ltd, Moore, R. & Centre for Sustainable Energy. 2008. *Fuel Poverty in London: Figures and tables illustrating the challenge of tackling fuel poverty* [Online]. London: Greater London Authority. Available: http://legacy.london.gov.uk/mayor/publications/2009/docs/fuel-poverty-jul09.pdf [Accessed 20 May 2011].

Aydinalp-Koksal, M. & Ugursal, V. I. 2008. Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector. *Applied Energy,* 85**,** 271-296.

Beaumont, J. R. & Inglis, K. 1989. Geodemographics in Practice - Developments in Britain and Europe. *Environment and Planning A,* 21**,** 587-604.

Bechtel, R. B. 1980. *What are post-occupancy evaluation : A Laymans guide to the POE for Housing. Draft final report,* Washington, DC, US Department of Housing and Urban Development.

Bennett, M. & Newborough, M. 2001. Auditing energy use in cities. *Energy Policy,* 29**,** 125-134.

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kuhnert, C. & West, G. B. 2007. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 7301-7306.

Bijmolt, T. H. A., Van Heerde, H. J. & Pieters, R. G. M. 2005. New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research,* 42**,** 141-156.

Bioregional. 2011. *OnePlanetLiving* [Online]. Available: http://www.oneplanetliving.org/index.html [Accessed 23 September 2011].

Boardman, B. 1988. Economic, social and technical considerations for fuel poverty policy. Brighton, UK: University of Sussex.

Boardman, B. 2005. *40% house,* Oxford, Environmental Change Insititute, University of Oxford.
Bond, S. & Insalaco, F. 2007. Area Classification of Super Output Areas and Datazones. Project Final Report. London: Office for National Statistics.
Booth, C. 1902. *Life and labour of the people in London,* London, Macmillan and Co.
Bordass, W. 2007. Energy Performance. *In:* Turrent, D. (ed.) *Sustainable Architecture.* London: RIBA Publications.
Box, G. E. P. & Cox, D. R. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* 26**,** 211-252.
BRE 1998. *The Government's standard assessment procedure for energy rating of dwellings,* Garston, Watford, UK, BRE.
BRE 2001. *The Government's standard assessment procedure for energy rating of dwellings,* Watford, BRE.
BRE 2005a. Energy Use in Homes: Fuel Consumption. London: HMSO.
BRE 2005b. Standard Assessment Procedure 2005. Garston, UK: BRE.
BRE 2009. A Consultation on proposed changes to the Government's Standard Assessment Procedure: Background and consulation questions. *In:* Change, D. o. E. a. C. (ed.). Garston, UK: BRE Group.
BRE 2010. The Government's Standard Assessment Procedure for Energy Rating of Dwellings: incorporating RdSAP 2009. 2009 ed. Garston: BRE Group.
British Council for Offices 2007. *British Council for Offices guide to post-occupancy evaluation.,* London, British Council for Offices.
British Gas. 2011. *Our tariffs from A-Z* [Online]. Available: http://www.britishgas.co.uk/products-and-services/energy/our-tariffs/tariff-A-Z.html [Accessed 15 March 2011].
Brown, R., Parker, D. & Homan, G. 2007. Appliances, Lighting, Electronics, and Miscellaneous Equipment Electricity Use in New Homes. Berkeley: Ernest Orlando Lawrence Berkeley National Laboratory.
Bruhns, H., Hamilton, I., Lowe, R., Steadman, J. P. & Summerfield, A. 2011. Buildings and Energy Data Frameworks. London: UCL Energy Institute.
Building Research Energy Conservation Support Unit 2007. Monitored Domestic Energy Use Data Archive, 1973-1983. SN: 2210 ed. Colchester, Essex: UK Data Archive.
Building Research Establishment 1956. Domestic Heating - Estimation of Seasonal Heat Requirement and Fuel Consumption in Houses. *BRE Digest.* Garston, Watford, UK: Building Research Establishment.
Building Research Establishment 1976. Heat Losses from Dwellings. *BRE Digest.* Garston, Watford, UK: Building Research Establishment.
Cambridge Architectural Research Ltd. 2011. *Four decades of housing energy data* [Online]. Cambridge, UK. Available: http://www.carltd.com/feature44.htm [Accessed 30 April 2011].
Carey, P. L. L. M. 2009. *Data protection : a practical guide to UK and EU law,* Oxford, Oxford University Press.
CECED. 2011. *The New EU Energy Label* [Online]. Available: http://www.newenergylabel.com/index.php/uk/home/ [Accessed 10 August 2011].
Census Advisory Working Group 1999. Population Definitions for the 2001 Census. *In:* Statistics, O. f. N. (ed.). London.
Chambers, J. M. 1983. *Graphical methods for data analysis,* Belmont, Wadsworth ; Boston : Duxbury.
Chapman, J., Lowe, R. & Everett, R. 1985a. *The Pennyland Project : this is the final report of 177 low-energy houses at Pennyland, Milton Keynes, monitored by the Open University Energy Research Group (ERG), for the Milton Keynes Development Corporation (MKDC), under contract to the Energy Technology Support Unit (ETSU) at Harwell,* Harwell, ETSU.

Chapman, J., Lowe, R. & Everett, R. 1985b. *The Pennyland Project : this is the final report of 177 low-energy houses at Pennyland, Milton Keynes, monitored by the Open University Energy Research Group (ERG), for the Milton Keynes Development Corporation (MKDC), under contract to the Energy Technology Support Unit (ETSU) at Harwell,* Harwell, ETSU.
Chapman, P. F. 1990. The Milton Keynes Energy Cost Index. *Energy and Buildings,* 14**,** 83-101.
Chishimba, L., Morris, J., Dudgon, M. & Beckett, P. 2009. Use of Experian MOSAIC to explore variability in process and outcomes in lung cancer patients. *Thorax,* 64**,** P170.
CIBSE 2006. *Environmental design : CIBSE guide A,* London, Chartered Institution of Building Services Engineers.
Clark, W. A. V., Deurloo, M. C. & Dieleman, F. M. 2006. Residential mobility and neighbourhood outcomes. *Housing Studies,* 21**,** 323-342.
Cohen, J. 1968. Multiple Regression as a General Data-Analytic System. *Psychological Bulletin,* 70**,** 426-&.
Coker, P. 2009. Presentation. Modelling supply and demand to explore the variability characteristics of low carbon energy alternatives. London, University College London.
Conover, W. J. 1999. *Practical nonparametric statistics,* New York ; Chichester, John Wiley.
Cooke, T. J. 2009. Selection Bias. *In:* Thrift, N. J. & Kitchin, R. (eds.) *International encyclopedia of human geography.* Amsterdam ; London: Elsevier Science.
Cooper, S. A. 1981. *Fuel poverty in the United Kingdom,* London, Policy Studies Institute for the Commission of the European Communities.
Cornelissen, G., Pandelaere, M. & Warlop, L. 2006. Cueing common ecological behaviors to increase environmental attitudes. *Persuasive Technology,* 3962**,** 39-44.
Courtney, R. 1997. Building research establishment - past, present and future - BRE's changing roles over time, the developments leading to its private status and BRE's future plans. *Building Research and Information,* 25**,** 285-291.
Cozby, P. C. 1997. *Methods in Behavioral Research,* Mountain View, Calif., USA, Mayfield.
Crawley, D. B., Hand, J. W., Kurnmert, M. & Griffith, B. T. 2008. Contrasting the capabilities of building energy performance simulation programs. *Building and Environment,* 43**,** 661-673.
Crosbie, T. 2006. Household energy studies: The gap between theory and method. *Energy and Environment,* 17**,** 735-753.
Crosbie, T. & Baker, K. 2010. Energy-efficiency interventions in housing: learning from the inhabitants. *Building Research and Information,* 38**,** 70-79.
Crosby, T. 2006. Household energy studies: the gap between theory and method. *Energy and Environment* 17**,** 735-753.
Darnton, A. 2006. Shaping the Energy-Related Behaviour of Future Consumers. London: Energy Saving Trust.
Davies, W. K. D. 1978. Charles Booth and the Measurement of Urban Social Character. *Area,* 10**,** 290-296.
Day, T., Jones, P. & Ogumka, P. 2007. Review of the impact of the energy policies in the London Plan on Applications referred to the Mayor (Phase 2). London: London South Bank University.
Department for Communities and Local Government 2006a. *The Building Regulations 2000. Conservation of fuel and power. Approved document L1A,* London, NBS.
Department for Communities and Local Government 2006b. Fuel Poverty Dataset Documentation. *UK Data Archive Study Number 6106 - English House Condition Survey, 2006.* London.
Department for Communities and Local Government 2007a. Building a Green Future: Policy Statement. London: HMSO.
Department for Communities and Local Government 2007b. *Regulatory impact assessment Energy Performance of Buildings Directive, articles 7-10 : the Energy Performance of Buildings*

(Certificates and Inspections) (England and Wales) Regulations 2007, London, Dept. for Communities and Local Government.
Department for Communities and Local Government 2008a. *The code for sustainable homes : setting the standard in sustainability for new homes,* London, Department for Communities and Local Government.
Department for Communities and Local Government 2008b. Notice of Approval of the methodology of calculation of the energy performance of buildings in England and Wales. London: Department for Communities and Local Government.
Department for Communities and Local Government 2009. Planning policy statement : Eco-towns, a supplement to planning policy statement 1. London: Department for Communities and Local Governmen.
Department for Communities and Local Government. 2010a. *English House Condition Survey (EHCS)* [Online]. London: Department for Communities and Local Government. Available: http://www.communities.gov.uk/housing/housingresearch/housingsurveys/englishhousecondition/ [Accessed 9 March 2010].
Department for Communities and Local Government 2010b. Planning Policy Statement 3: Housing. *In:* Government, D. f. C. a. L. (ed.). London.
Department for Communities and Local Government. 2011a. *Dwelling Stock Estimates: 2011, England* [Online]. Available: http://www.communities.gov.uk/documents/statistics/pdf/2039750.pdf [Accessed 30 April 2011].
Department for Communities and Local Government. 2011b. *English Housing Survey (EHS)* [Online]. London: Department for Communities and Local Government. Available: http://www.communities.gov.uk/housing/housingresearch/housingsurveys/englishhousingsurvey/ [Accessed 15 March 2011].
Department for Energy and Climate Change 2010a. Average annual domestic electricity bills for selected towns and cities in the UK and average unit costs. *In:* qep223.xls (ed.). London.
Department for Energy and Climate Change 2010b. Guidance note for the DECC MLSOA/IGZ and LLSOA electricity and gas consumption data London: Department for Energy and Climate Change.
Department for Environment Food and Rural Affairs 2009. Saving energy through better products and appliances: A report on analysis, aims and indicative standards for energy efficient products 2009 - 2030. London: Defra.
Department for the Environment Food and Rural Affairs. 2007. *Report, Questionnaire and Data Tables Following Survey of Public Attitudes and Behaviours Toward the Environment* [Online]. London: DEFRA. Available: http://www.defra.gov.uk/evidence/statistics/environment/pubatt/download/pubattsum2007.pdf [Accessed].
Department of Energy 1978. Energy policy : a consultative document. London: H.M.S.O.
Department of Energy and Climate Change 2009a. *Digest of United Kingdom energy statistics*, London : HMSO.
Department of Energy and Climate Change 2009b. *The UK low carbon transition plan : national strategy for climate and energy,* London, Stationery Office.
Department of Energy and Climate Change. 2010. *Response to Local Authority consultation on sub-regional fuel poverty data* [Online]. Available: http://www.decc.gov.uk/assets/decc/Statistics/fuelpoverty/933-response-la-conssubregional-fuelpov-stats.pdf [Accessed].
Department of Energy and Climate Change 2011. Great Britain's housing energy fact file. London: Department of Energy and Climate Change.

Department of Energy and Climate Change. 2012. *Smart Metering Implementation Programme: Data access and privacy consultation draft* [Online]. Available: http://www.decc.gov.uk/assets/decc/11/consultation/smart-metering-imp-prog/4933-data-access-privacy-con-doc-smart-meter.pdf [Accessed 12 April 2012].
Department of Energy and Climate Change and BRE. 2010. *Fuel Poverty Sub-Regional Statistics* [Online]. Available: http://www.decc.gov.uk/en/content/cms/statistics/fuelpov_stats/regional/regional.aspx [Accessed].
Department of the Environment 1995. Approved Document Part L. London: HMSO.
Department of the Environment 1996. Home Energy Conservation Act 1995. London: HMSO.
Department of the Environment Transport and the Regions 2000a. *By design : urban design in the planning system; towards better practice*, Thomas Telford Publishing.
Department of the Environment Transport and the Regions 2000b. English House Condition Survey 1996 User Guide. London: Department of the Environment Transport and the Regions.
Department of the Environment Transport and the Regions 2002. Methodological Review of the Survey of English Housing. London: Department of the Environment Transport and the Regions.
DETR 2000. *By design : urban design in the planning system; towards better practice,* London, Thomas Telford Publishing.
Dickson, C. M., Dunster, J. E., Lafferty, S. Z. & Shorrock, L. D. 1996. BREDEM: Testing monthly and seasonal versions against measurements and against detailed simulation models. *Building Services Engineering Research and Technology,* 17**,** 135-140.
DoE 1990. Approved Document L1 - Conservation of fuel and power. London: Department of the Environment and the Welsh Office.
Doran, S. M. & Anderson, B. R. 1995. *BREDEM development for conservatories and passive solar houses,* Garston, Watford, UK, Building Research Establishment.
Draper, N. R. & Smith, H. 1981. *Applied regression analysis,* New York ; Chichester, Wiley.
Economic and Social Data Service. 2011. *About the Economic and Social Data Service* [Online]. Available: http://www.esds.ac.uk/about/about.asp [Accessed 21 March 2011].
Ekins, P. & Dresner, S. 2006. Modelling Future Household Energy Use and Fuel Poverty: A Scoping Study. London: Policy Studies Institute.
Electric Power Research Institute 1989. Residential end-use energy consumption: a survey of conditional demand analysis. Palo Alto, Calif., USA.
Elexon. 2010. *What Does ELEXON Do?* [Online]. Available: http://www.elexon.co.uk/aboutelexon/whatelexondoes.aspx [Accessed 4 October 2010].
Energy Advisory Services 1996. BREDEM-12: Supporting Evidence. Unpublished.
Energy Information Administration 2001. Residential Energy Consumption Survey. Washington, D.C.
Energy Saving Trust 2004. Monitoring Energy Savings achieved from Insulation Measures installed in Gas-heated Homes in SoP3 and EEC Schemes. *Energy Savings Monitoring Report to Defra.* London: Department of Energy and Climate Change.
Energy Saving Trust 2007a. The Ampere Strikes Back. London: Energy Saving Trust.
Energy Saving Trust. 2007b. *Planet Pays the Price for Home Entertainment* [Online]. London: Energy Saving Trust. Available: http://www.energysavingtrust.org.uk/Resources/Features/Features-archive/Planet-pays-the-price-of-home-entertainment [Accessed 26 May 2009].
Energy Saving Trust 2008. Homes Energy Efficiency Database (HEED): An overview. London: Energy Saving Trust.
Energy Saving Trust 2009a. Area-Based Approaches: A Good Practice Guide. London.

Energy Saving Trust 2009b. Final Report: In-situ monitoring of efficiencies of condensing boilers and use of secondary heating. *In:* CRE, G. a. (ed.). London Department of Energy and Climate Change.
Energy Saving Trust 2011. The elephant in the living room: How our appliances and gadgets are trampling the green dream. London: Energy Saving Trust.
Environmental Change Institute 1995. DECADE: Domestic equipment and carbon dioxide emissions, Second Year Report 1995. Oxford: Oxford University.
Espey, M., Espey, J. & Shaw, W. D. 1997. Price elasticity of residential demand for water: A meta-analysis. *Water Resources Research,* 33**,** 1369-1374.
ESRC Census Programme. 2006. *Aggregate Statistics* [Online]. Available: https://www.census.ac.uk/guides/Area_stats.aspx [Accessed 25 July 2011].
EURECO 2002. Demand-Side Management: End-use metering campaign in 400 households of the European Community Assessment of the Potential Electricity Savings. Brussels: Commission of the European Communities.
European Central Bank 2003. Structural factors in the EU housing markets. Frankfurt am Main: European Central Bank.
European Commission. 2010. *Directive 2010/31/EU: Environmental Performance of Buildings Directive* [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:153:0013:0035:EN:PDF [Accessed 11 August 2011].
European Committee for Standardization 1992. Residential buildings - Energy requirements for heating - Calculation method. Brussels.
European Parliament and European Council 2002. Directive 2002/91/EC of the European Parliament and of the Council of 16 December 2002 on the energy performance of buildings. Brussels: European Community.
Everett, R., Horton, A., Doggart, J. & Willoughby, J. 1985. *Linford low energy houses,* Harwell, Department of Energy Energy Technology Support Unit.
Everitt, B. 2011. *Cluster analysis,* Oxford, Wiley-Blackwell.
Experian UK. 2009. *Mosaic E-Flipchart* [Online]. London: Experian UK. Available: http://cdu.mimas.ac.uk/experian/mosaic.hlp [Accessed 15 March 2009 2009].
Experian UK. 2010. *About Experian* [Online]. London: Experian. Available: http://www.experian.co.uk/www/pages/about_us/index.html [Accessed 10 February 2010 2010].
Fawcett, T. 2000. *Lower carbon futures for European households,* Oxford, Environmental Change Institute.
Feijten, P. & van Ham, M. 2009. Neighbourhood Change ... Reason to Leave? *Urban Studies,* 46**,** 2103-2122.
Field, A. P. & Miles, J. 2010a. *Discovering statistics using SAS,* London, SAGE.
Field, A. P. & Miles, J. 2010b. *Discovering statistics using SAS : (and sex and drugs and rock 'n' roll),* London, SAGE.
Firth, S., Lomas, K., Wright, A. & Wall, R. 2008. Identifying trends in the use of domestic appliances from household electricity consumption measurements. *Energy and Buildings,* 40**,** 926-936.
Firth, S. K., Lomas, K. J. & Rees, S. J. 2010. A simple model of PV system performance and its use in fault detection. *Solar Energy,* 84**,** 624-635.
Fisher, R. A. S. 1973. *Statistical methods for research workers ... Fourteenth edition, revised and enlarged*, New York: Hafner Publishing Co.
Fox, J. 2002. *Robust Regression: Appendix to An R and S-Plus Companion to Applied Regression* [Online]. Available: http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-robust-regression.pdf [Accessed 4 August 2011].

Freedman, D. A. 1999. Ecological Inference and the Ecological Fallacy. Berkeley: International Encyclopedia of the Social & Behavioral Sciences.
Gaze, C. 2008. *Applying the code for sustainable homes on the BRE Innovation Park,* Garston, Watford, UK, BRE.
Geoinformaton Group. 2012. *About National Building Class* [Online]. Available: http://www.geoinformationgroup.co.uk/products/building-class [Accessed 11 July 2012].
Global Cool. 2011. *Do It In Public* [Online]. Available: http://www.globalcool.org/tag/do-it-in-public [Accessed 10 August 2011].
Great Britain Parliament 1984. The Building Act 1984. London: HMSO.
Great Britain Parliament 1998. Data Protection Act 1998. London: HMSO.
Great Britain Parliament 2000. Warm Homes and Energy Conservation Act 2000. London: HMSO.
Greene, W. H. 2008. *Econometric analysis,* Upper Saddle River, N.J., Pearson Prentice Hall.
Groom, R. 2009. UKMap - more competition for Ordnance Survey? *Geomatics world (Lemmer),* 17**,** 28-29.
Group for Efficient Appliances 1995. Washing Machines, Driers, and Dishwashers: Final Report. Copenhagen: Danish Energy Agency.
Guerin, G. A., Yust, B. L. & Coopet, J. G. 2000. Occupant predictors of household energy behavior and consumption change as found in energy studies since 1975. *Family and Consumer Sciences Research Journal* 29**,** 48-80.
Guler, B., Fung, A. S., Aydinalp, M. & Ugursal, V. I. 2001. Impact of energy efficiency upgrade retrofits on the residential energy consumption in Canada. *International Journal of Energy Research,* 25**,** 785-792.
Guy, S. 2006. Designing urban knowledge: competing perspectives on energy and buildings. *Environment and Planning C-Government and Policy,* 24**,** 645-659.
Guy, S. & Marvin, S. 1999. Understanding sustainable cities: Competing urban futures. *European Urban and Regional Studies,* 6**,** 268-275.
Haines, M. R. 1995. Socioeconomic differentials in infant and child-mortality during mortality decline - England and Wales, 1890-1911. *Population Studies,* 49**,** 297-315.
Haldane, R. B. H. V. 1918. *Report of the Machinery of government committee*, [S.l.] : H.M.S.O., 1925 (1918).
Hall, P. G. & Ward, C. 1998. *Sociable cities : the legacy of Ebenezer Howard,* Chichester, J. Wiley.
Harr, S. & Reed, C. 1996. Coming Home: A postscript on postmodernism. *In:* Reed, C. (ed.) *Not at home : the suppression of domesticity in modern art and architecture.* London: Thames & Hudson.
Harris, R. 1997. 'The nature of cities' and urban geography in the last half century. *Urban Geography,* 18**,** 15-35.
Harris, R., Sleight, P. & Webber, R. 2005. *Geodemographics, GIS and neighbourhood targeting,* Chichester, John Wiley & Sons.
Hartley, H. O. & Rao, J. N. K. 1967. Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model. *Biometrika,* 54.
Hartman, R. S. 1988. Self-Selection Bias in the Evolution of Voluntary Energy Conservation Programs *The Review of Economics and Statistics,* 70**,** 448-458.
Hassell, C. & Olivier, D. 2005. *The Green Electricity Illusion* [Online]. ech2o. Available: http://www.ech2o.co.uk/downloads/GEI.pdf [Accessed 11 May 2011].
Henderson, G. & Shorrock, L. D. 1986a. BREDEM - THE BRE DOMESTIC ENERGY MODEL: TESTING THE PREDICTIONS OF A TWO-ZONE VERSION. *Building services engineering research & technology,* 7**,** 87-91.

Henderson, G. & Shorrock, L. D. 1986b. BREDEM - the BRE Domestic Energy Model: testing the predictions of a two-zone version. *Building Services Engineering Research and Technology,* 7**,** 87-91.
Henderson, J. 2009. Review of auxiliary energy use and the internal heat gains assumptions in SAP. Garston, Watford, UK: BRE Group.
Henwood, M. 1997. Fuel poverty, energy efficiency and health: a report to the EAGA Charitable Trust. Keswick, UK: EAGA-CT.
Herbert, D. 1972. *Urban geography: a social perspective*, Newton Abbot: David and Charles.
Hinnells, M., Boardman, B., Layberry, R., Darby, S. & Killip, G. 2007. The UK Housing Stock 2005 to 2050: Assumptions used in Scenarios and Sensitivity Analysis in UKDCM2. Oxford: Enviromental Change Institute.
Hinnells, M. J., Lane, K. B., Small, E. C., Boardman, B. & Middleton, N. 1995. Policy Options: Energy and savings from intervention. Oxford, UK: Environmental Change Unit.
HM Government 2010. 2050 Pathways Analysis. *In:* Change, D. o. E. a. C. (ed.). London.
HMSO 1998. Data Protection Act 1998. *In:* Office, H. (ed.). London.
Homes and Communities Agency 2007. *Urban design compendium 2: Delivering Quality Places,* London, English Partnerships.
Hong, S. H., Oreszczyn, T., Ridley, I. & Grp, W. F. S. 2006a. The impact of energy efficient refurbishment on the space heating fuel consumption in English dwellings. *Energy and Buildings,* 38**,** 1171-1181.
Hong, S. H., Oreszczyn, T., Ridley, I. & Warm Front Study Group 2006b. The impact of energy efficient refurbishment on the space heating fuel consumption in English dwellings. *Energy and Buildings,* 38**,** 1171-1181.
Howick, R. 2004. Building neighbourhood classifications - data sources and their geographic integration. *ESRC Transdisciplinary/Research Methods Seminar Series.* London.
Huber, P. 1964. Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics,* 35.
Humphrey, S., Loh, J. & Goldfinger, S. 2008. *Living planet report 2008,* Gland, Switzerland, WWF.
Imperial College and Energy Networks Assocation 2010. Benefits of Advanced Smart Metering for Demand Response based Control of Distribution Networks. 2 ed. London: Imperial College.
Impetus Consulting Ltd 2006. Review of UK energy efficiency data needs. London: Energy Efficiency Partnership for Homes.
Inter-jurisdictional Regulatory Collaboration Committee. 2010. *Performance-Based Building Regulatory Systems: Principles and Experiences* [Online]. Available: http://www.irccbuildingregulations.org/pdf/A1163909.pdf [Accessed 24 January 2012].
International Energy Agency 2007. European and Canadian non-HVAC Electric and DHW Load Profiles for Use in Simulating the Performance of Residential Cogeneration Systems. *In:* Knight, I. & Ribberink, H. (eds.) *Annex 42 The Simulation of Building-Integrated Fuel Cell and Other Cogeneration Systems (COGEN-SIM).* Ottawa: CANMET Energy Technology Centre.
International Energy Agency 2009. *Gadgets and Gigawatts - Policies for Energy Efficient Electronics,* Paris.
International Initiatives for a Sustainable Built Environment. 2010. *SB Method 2010* [Online]. Available: http://www.iisbe.org/sbmethod-2010 [Accessed 20 July 2011].
International Organization for Standardisation 1989. Calculation of space heating requirements for residential buildings. Geneva: ISO.
International Union of Pure and Applied Chemistry 2010. *Mendeleev and his impact on the development of science*, IUPAC.
Jackson, C., Best, N. & Richardson, S. 2008. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society Series a-Statistics in Society,* 171**,** 159-178.

Jacobs, J. 1962. *The Death and Life of Great American Cities,* London, Jonathan Cape.
Jacobsen, H. K. 1998. Integrating the bottom-up and top-down approach to energy-economic modelling: the case of Denmark. *Energy Economics,* 20**,** 443-461.
James, N. & Desai, P. 2003. *One planet living in the Thames Gateway : a WWF-UK one million sustainable homes campaign report,* Godalming, WWF-UK.
Jelsma, J. 2002. Smart work package 4.2: Smart field test: experience of users and technical aspects. Petten, Netherlands: Energy Research Centre of the Netherlands.
Jevons, W. S. 1865. *The coal question : an inquiry concerning the progress of the nation and the probable exhaustion of our coal-mines,* London, Macmillan.
Johnston, D., Lowe, R. & Bell, M. 2005. An exploration of the technical feasibility of achieving CO2 emission reductions in excess of 60% within the UK housing stock by the year 2050. *Energy Policy,* 33**,** 1643–1659.
Johnston, R. J. 1971. *Urban residential patterns: an introductory review*, London: Bell.
Jones, N. Year. The use of energy ratings in the UK. *In:* Climate technology and energy audit for improved energy efficiency, 26-28 September 2000 2000 Tallinn.
Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, A., Stevanovic, Z. & Djurovic-Petrovic, M. 2010. A review of bottom-up building stock models for energy consumption in the residential sector. *Building and Environment,* 45**,** 1683-1697.
Kaza, N. 2010. Understanding the spectrum of residential energy consumption: A quantile regression approach. *Energy Policy,* 38**,** 6574-6585.
Keating, K. M. 1989. Self-Selection - Are We Beating a Dead Horse. *Evaluation and Program Planning,* 12**,** 137-142.
Keirstead, J. 2006. Evaluating the applicability of integrated domestic energy consumption frameworks in the UK. *Energy Policy,* 34**,** 3065-3077.
King, G., Rosen, O. & Tanner, M. A. 2004. *Ecological inference : new methodological strategies,* Cambridge, UK ; New York, Cambridge University Press.
Kohler, N. & Hassler, U. 2002. The building stock as a research object. *Building Research and Information,* 30**,** 226-236.
Landmap Service. 2011. *Explore Building Classes for Some of the Major Urban Areas in the UK* [Online]. Available: http://landmap.mimas.ac.uk/index.php/Datasets/Feature/Building-Class/menu-id-100211.html [Accessed 21 September 2011].
Lavin, F. V., Dale, L., Hanemann, M. & Moezzi, M. 2011. The impact of price on residential demand for electricity and natural gas. *Climatic Change,* 109**,** 171-189.
Lebot, B., Lenci, O. & Waide, P. 1995. Measuring electricity consumption by end-use : lessons learned from a monitoring project in the residential sector. *ECEEE summer study, The energy efficiency challenge for Europe.* Mandelieu , France.
Lenzen, M., Dey, C. & Foran, B. 2004. Energy requirements of Sydney households. *Ecological Economics,* 49**,** 375-399.
Likert, R. 1932. *A Technique for the Measurement of Attitudes,* New York, Harper.
Lorimer, S. 2010. Potential for a neighbourhood income-based domestic energy model for ordinary electricity use in England. *Central Europe Sustainable Building Conference 10.* Prague.
Lowe, R. 2007. Technical options and strategies for decarbonizing UK housing. *Building Research and Information,* 35**,** 412-425.
Lowe, R. 2010. *RE: Meeting Thursday 18 March.* Type to Lorimer, S.
Lowe, R., Chiu, L. F. & Bell, M. 2009. LowCarb4Real. *The Complete UrbanBuzz.* London: University College London.
Lutters, W. G. & Ackerman, M. S. 1996. An Introduction to the Chicago School of Sociology. Baltimore: University of Maryland-Baltimore County.
Lutzenhiser, L. 1992. A Cultural Model of Household Energy-Consumption. *Energy,* 17**,** 47-60.

Lutzenhiser, L., Moezzi, M., Hungerford, D. & Friedman, R. Year. Sticky Points in Modeling Household Energy Consumption. *In:* ACEEE Summer Study on Energy Efficiency in Buildings: The Climate for Efficiency is Now, 2010 Pacific Grove, Calif., USA. American Council for an Energy-Efficient Economy Home.
Macmillian, S. & Kohler, N. 2004. Modelling energy use in the global building stock: a pilot survey to identify available data sources. Cambridge, UK: University of Cambridge.
Market Transformation Programme 2008. BNCK01: Assumptions underlying the energy projections of cooking appliances. Didcot, UK: Defra.
Market Transformation Programme 2010. BNXS01: Carbon Dioxide Emission Factors for UK Energy Use. London: Department of Energy and Climate Change.
Martin, D. 2001. Geography for the 2001 Census in England and Wales. London: Office of National Statistics.
Mayor of London 2009. London Housing Design Guide: Draft for Consultation. London: Greater London Authority.
McIntyre, T. 5 April 2011 2011. *RE: Energy Follow Up Survey.* Type to Lorimer, S.
McMichael, M. 2009. Is it about 'who you know'? The role of social capital in UK household energy consumption. *Behavior, Energy and Climate Change Conference.* Washington, D.C.
McPherson, M., Smith-Lovin, L. & Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology,* 27**,** 415-444.
Meacham, B., Bowen, R., Traw, J. & Moore, A. 2005. Performance-based building regulation: current situation and future needs. *Building Research and Information,* 33**,** 91-106.
Moore, R. 2008. Retrofitting the existing housing stock in the South East. London: Centre for Sustainable Energy / Association for the Conservation of Energy.
Myers, J. L., Well, A. & Lorch, R. F. 2010. *Research design and statistical analysis,* London, Routledge Academic.
Natarajan, S. & Levermore, G. J. 2007. Predicting future UK housing stock and carbon emissions. *Energy Policy,* 35**,** 5719-5727.
Neffendorf, H., Bruhns, H. & Harrison, A. 2009. Non-domestic Energy Efficiency Data Framework (NEED): Scoping Study Findings. London: Department of Energy and Climate Change.
Nesbakken, R. 1999. Price sensitivity of residential energy consumption in Norway. *Energy Economics,* 21**,** 493-515.
Nguyen, C. 2007. Histogram of Numeric Data Distribution from the UNIVARIATE Procedure. *NorthEast SAS Users Group (NESUG)2007.* Baltimore, Md., USA.
Npower. 2011. *Prices in your area* [Online]. Available: https://www.npower.com/at_home/Applications/QuoteAndSwitch/UnitPrices.aspx?workflow=UnitPrice [Accessed 15 March 2011].
O'Neill, B. C. & Chen, B. S. 2002. Demographic determinants of household energy use in the United States. *Population and Development Review,* 28**,** 53-88.
Office for National Statistics 2004. *Census 2001 : key statistics for urban areas in England and Wales : laid before Parliament pursuant to section 4(1) Census Act 1920,* London, TSO.
Office for National Statistics. 2005a. *Area classification for output areas* [Online]. London: Office for National Statistics. Available: http://www.statistics.gov.uk/about/methodology-by-theme/areaclassification/oa/default.asp [Accessed].
Office for National Statistics 2005b. Number of People Living in Households (UV51). London: Office for National Statistics.
Office for National Statistics 2005c. Rooms, Amenities, Central Heating and Lowest Floor Level (KS19). London: Office for National Statistics: Neighbourhood Statistics.
Office for National Statistics 2008a. Demographic information, household composition and relationships. 2 ed. London: Office for National Statistics.

Office for National Statistics 2008b. Guidance notes for the 2001 Area Classification of Super Output Areas and Data Zones. London: Office for National Statistics.
Office for National Statistics 2010. Social Trends 40 ed. London: Office for National Statistics.
Office for National Statistics and Department for Environment Food and Rural Affairs 2010. Living Costs and Food Survey, 2008. Colchester, Essex: UK Data Archive.
Office of National Statistics 2005. Number of Rooms (UV57). London: Department for Communities and Local Government.
Office of the Deputy Prime Minister 2000. Approved Document Part L. London: HMSO.
Openshaw, S. 1984. Ecological Fallacies and the Analysis of Areal Census-Data. *Environment and Planning A,* 16**,** 17-31.
Openshaw, S., Cullingford, D. & Gillard, A. 1980. A Critique of the National Classifications of OPCS/PRAG. *The Town Planning Review,* 51**,** 421-439.
Palmer, J. Year. Cutting carbon emissions from UK homes: Are we on track? *In:* Energy Consumption From Dwellings: Do We Understand It?, 11 November 2010 2010 Cambridge, UK. GreenBRIDGE Seminar Series.
Pank, W., Girardet, H. & Cox, G. 2002. *Tall buildings and sustainability : report,* London, Corporation of London.
Parekh, A. Year. Development of archetypes of building characteristics libraries for simplified energy use evaluation of houses. *In:* IBPSA, ninth international conference, 2005 Montreal. 921–8.
Parliamentary Office of Science and Technology. 2005. *POSTnote 249: Household Energy Efficiency* [Online]. Available: http://www.parliament.uk/documents/post/postpn249.pdf [Accessed 10 August 2008].
Parti, M. & Parti, C. 1980. The Total and Appliance-Specific Conditional Demand for Electricity in the Household Sector. *Bell Journal of Economics,* 11**,** 309-321.
Pears, A. & Versluis, P. 1993. *Scenarios for alternative energy in Western Australia : a report for the renewable energy advisory council,* Perth, Renewable Energy Advisory Council.
Pearson, L. F. 1981. Hull Low Energy Housing Project : Social Survey, 1981. SN: 1589 ed. Colchester, Essex: UK Data Archive
Peter Warm 1999. Hot Water Surveys. Oxford, UK: Environmental Change Institute, Oxford University.
Pinker, S. 1997. *How the mind works,* London, Allen Lane, 1998.
Prior, J. J. & Williams, C. 2008. *Sustainability through planning : local authority use of BREEAM, EcoHomes and the Code for Sustainable Homes,* Bracknell, BRE Press.
Raudenbush, S. W. & Bryk, A. S. 2002. *Hierarchical linear models : applications and data analysis methods,* Thousand Oaks, CA ; London, Sage Publications.
Rees, P. H. 1979. *Residential patterns in American cities: 1960*
University of Chicago, Department of Geography, Research Paper.
RIBA / CIBSE. 2009. *Carbon Buzz: a RIBA / CIBSE Platform* [Online]. London: University College London. Available: http://www.bre.co.uk/carbonbuzz/ [Accessed 20 March 2011].
Ritchie, A. M. & Thomas, R. 2009. *Sustainable urban design : an environmental approach,* London, Taylor & Francis.
Robinson, W. S. 1950. Ecological Correlations and the Behavior of Individuals. *American Sociological Review,* 15**,** 351-357.
Robinson, W. S. 2009. Ecological Correlations and the Behavior of Individuals. *International Journal of Epidemiology,* 38**,** 337-341.
Robson, B. T. 1971. *Urban analysis : a study of city structure with special reference to Sunderland,* [S.l.], [s.n.].
Romig, F. & Leach, G. 1977. *Energy conservation in UK dwellings : Domestic sector survey & insulation,* [S.l.], International Institute for Environment and Development.

Rose, D. F. & Pevalin, D. J. 2003. *A researcher's guide to the national statistics socio-economic classification,* London, SAGE.
Sanders, H. D. 1992. The Khazzoom-Brookes postulate and neoclassical growth. *Energy Policy,* 13**,** 131-148.
SAS Institute 2011. *Base SAS(R) 9.2 Procedures Guide: Statistical Procedures,* Cary, North Carolina, USA, SAS Institute.
Schuessler, A. A. 1999. Ecological inference. *Proceedings of the National Academy of Sciences of the United States of America,* 96**,** 10578-10581.
Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J. & Griskevicius, V. 2007. The constructive, destructive and reconstructive power of social norms. *Psychological Science,* 18**,** 429-434.
Scottish Executive 2003. Building (Scotland) Act 2003. London: HMSO.
Scottish Government Social Research 2009. Modelling greenhouse gas emissions from Scottish housing: Final report. Edinburgh: Queens Printers of Scotland.
Sedlacek, S. & Maier, G. 2010. Do green building councils make sense? - An economic analysis. *CESB 10 - Central Europe towards sustainable building.* Prague: Grada Publishing for Department of Building Structures and CIDEAS Research Centre.
Sefton, T. & Chesshire, J. 2005. Peer review of the methodology for calculating the number of households in fuel poverty in England: Final report to DTI and DEFRA. *In:* Industry, D. f. T. a. & Department for the Environment, F., and Rural Affairs (eds.). London.
Shalizi, C. R. 2011. Scaling and Hierarchy in Urban Economies. *Proceedings of the National Academy of Sciences of the United States of America*.
Shipworth, M., Firth, S.K., Gentry, M.I., Wright, A.J., Shipworth, D.T., Lomas, K.J. 2010. Central heating thermostat settings and timing: building demographics. *Building Research and Information,* 38**,** 50-69.
Shorrock, L. 5 February 2010 2010. *RE: RE: History of BREDEM.* Type to Lorimer, S.
Shorrock, L. D. & Anderson, B. R. 1995. *A guide to the development of BREDEM,* Garston, Watford, UK, Building Research Establishment.
Shorrock, L. D. & Dunster, J. E. 1997. The physically-based model BREHOMES and its use in deriving scenarios for the energy use and carbon dioxide emissions of the UK housing stock. *Energy Policy,* 25**,** 1027-1038.
Shorrock, L. D., Dunster, J. E., Seale, C. F., Eppel, H. & Lomas, K. 1994. Testing BREDEM-8 against measured consumption data and against simulation models. *Building environmental performance : facing the future.*
Shorrock, L. D., Henderson, J. & Utley, J. I. 2005. *Reducing carbon emissions from the UK housing stock,* Garston, BRE Bookshop.
Shorrock, L. D., Macmillan, S., Clark, J. & Moore, G. 1991. BREDEM-8, a monthly calculation method for energy use in dwellings: Testing and development. *Building environmental performance '91.*
Shorrock, L. D. & Utley, J. I. 2008. *Domestic energy fact file 2008,* Garston, BREbookshop.
Simey, T. S. 1960. *Charles Booth Social Scientist,* [S.l.], Oxford University Press.
Simpson, L. & Brown, M. 2008. Census fieldwork in the UK: the bedrock for a decade of social analysis. *Environment and Planning A,* 40**,** 2132-2148.
Singer, J. 1998. Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioural Statistics,* 24**,** 323-355.
Singer, J. D. & Willett, J. B. 2003. *Applied longitudinal data analysis : modeling change and event occurrence,* Oxford, Oxford University Press.
Smith, A. Z. P. 2011. *RE: Personal communication.* Type to Lorimer, S.
Sørensen, B. 2011. *Renewable energy : physics, engineering, environmental impacts, economics & planning,* London, Academic.

Sorrell, S. 2009. Jevons' Paradox revisited: The evidence for backfire from improved energy efficiency. *Energy Policy,* 37**,** 1456-1469.
Sorrell, S., Dimitropoulos, J. & Sommerville, M. 2009. Empirical estimates of the direct rebound effect: A review. *Energy Policy,* 37**,** 1356-1371.
Statistics Commission 2007. *Counting on success : the 2011 census - managing the risks : November 2007,* London, Statistics Commission.
Steele, F. 2011. *Introduction to Multilevel Modelling* [Online]. Bristol: University of Bristol. Available: http://www.cmm.bris.ac.uk/lemma/mod/resource/view.php?id=193 [Accessed 21 May 2011].
Steemers, K. & Cheng, V. Year. Building Energy Demand and the Building Stock - ReVISIONS, WP10. *In:* ReVISIONS International Symposium, 6 September 2010 2010 Cambridge, UK.
Stern, P. C. 1986. Blind Spots in Policy Analysis - What Economics Doesnt Say About Energy Use. *Journal of Policy Analysis and Management,* 5**,** 200-227.
Stern, P. C. 2000. Toward a coherent theory of environmentally significant behavior. *Journal of Social Issues,* 56**,** 407-424.
Summerfield, A. J., Lowe, R. J., Bruhns, H. R., Caeiro, J. A., Steadman, J. P. & Oreszczyn, T. 2007. Milton Keynes Energy Park revisited: Changes in internal temperatures and energy usage. *Energy and Buildings,* 39**,** 783-791.
Summerfield, A. J., Lowe, R. J., Firth, S. K., Wall, R. & Oreszczyn, T. 2006. Carbon emissions and the case for joined-up research: Adding value to household and building energy datasets. *COBRA 2006: the construction and building research conference of the Royal Institution of Chartered Surveyors.* Royal Institute of Chartered Surveyors.
Summerfield, A. J., Lowe, R. J. & Oreszczyn, T. 2010. Two models for benchmarking UK domestic delivered energy. *Building Research and Information,* 38**,** 12-24.
Swan, L., Ugursal, V. I. & Beausoleil-Morrison, I. 2008. A new hybrid end-use energy and emissions model of the Canadian housing stock. *First International Conference on Building Energy and Environment, Proceedings Vols 1-3***,** 1843-1850
2365.
Swan, L. G. & Ugursal, V. I. 2009. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable & Sustainable Energy Reviews,* 13**,** 1819-1835.
Swan, L. G. & Urgusal, V. I. 2009. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews,* 13**,** 1819-1835.
Thomas, R. & Fordham, M. 2003. *Sustainable urban design : an environmental approach,* London, Spon Press.
Thornley, A. 1991. *Urban planning under Thatcherism : the challenge of the market,* London, Routledge.
Timms, D. W. G. 1971. *The urban mosaic: towards a theory of residential differentiation*, London: Cambridge University Press.
Tsao, J. Y., Saunders, H. D., Creighton, J. R., Coltrin, M. E. & Simmons, J. A. 2010. Solid-state lighting: an energy-economics perspective. *Journal of Physics D-Applied Physics,* 43.
U.S. Energy Information Administration 1997. End-Use Estimation Methodology. *In:* Administration, U. S. E. I. (ed.). Washington, DC.
U.S. Energy Information Administration 2011. 2009 Residential Energy Consumption Survey. *In:* http://www.eia.doe.gov/consumption/residential/reports/images/mainheatingsrc.xls (ed.). Washington, D.C.: U.S. Energy Information Administration.
UCLA. 2011. *SAS Annotated Output: Regression Analysis* [Online]. Los Angeles: UCL Academic Technology Services. Available: http://www.ats.ucla.edu/stat/sas/output/reg.htm [Accessed 27 June 2011].

Uglow, C. E. 1981. The calculation of energy use in dwellings. *Building Services Engineering Research and Technology,* 2**,** 1-14.
Uglow, C. E. 1982. Energy use in dwellings: An excercise to investigate the validity of a simple calculation method. *Building Services Engineering Research and Technology,* 3**,** 35-39.
UK Energy Research Centre. 2008. *Milton Keynes Energy Park Dwellings 1990* [Online]. London: UK Energy Research Centre. [Accessed 3 February 2010].
US Green Building Council. 2008. *LEED for Homes* [Online]. Available: http://www.usgbc.org/DisplayPage.aspx?CMSPageID=147 [Accessed 30 July 2011].
Valuation Office Agency. 2011. *Age Codes* [Online]. Available: http://www.voa.gov.uk/corporate/Publications/DwellingHouseCodingGuide/ageCodes.html [Accessed 20 July 2011].
van Ham, M. & Clark, W. A. V. 2009. Neighbourhood mobility in context: household moves and changing neighbourhoods in the Netherlands. *Environment and Planning A,* 41**,** 1442-1459.
Vickers, D. 2005. Methodology paper. *2001 Census Area Classification of Output Areas.* London: Office of National Statistics.
Vickers, D. 2006. *Multi-Level Integrated Classifications Based on the 2001 Census.* PhD, University of Leeds.
Vickers, D. & Rees, P. 2007. Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society Series a-Statistics in Society,* 170**,** 379-403.
Wakefield, J. 2004. Ecological inference for 2x2 tables. *Journal of the Royal Statistical Society Series a-Statistics in Society,* 167**,** 385-426.
Wall, R. & Crosbie, T. 2009. Potential for reducing electricity demand for lighting in households: An exploratory socio-technical study. *Energy Policy,* 37**,** 1021-1031.
Warwickshire Observatory. 2011. *Warwickshire Observatory Briefing Note: 2008 Fuel Poverty Statistics* [Online]. Warwick: Warwickshire County Council. Available: http://warksobservatory.files.wordpress.com/2011/04/briefing-note-fuel-poverty.pdf [Accessed 20 May 2011].
Watson, D. 1979. *Energy conservation through building design,* New York ; London, McGraw-Hill.
Webber, C. J. 1989. Using multiple data sources to build an area classification system%3A Operational problems encountered by MOSAIC. *Journal of the Market Research Society,* 31**,** 103-109.
Webber, R. & Craig, J. 1976. Which local authorities are alike? *Population Trends,* 5.
Webber, R. J. 1978. Making the Most of the Census for Strategic Analysis. *The Town Planning Review,* 49**,** 274-284.
Welz, T., Hischier, R. & Hilty, L. M. 2011. Environmental impacts of lighting technologies - Life cycle assessment and sensitivity analysis. *Environmental Impact Assessment Review,* 31**,** 334-343.
Westergren, K. E., Hogberg, H. & Norlen, U. 1999. Monitoring energy consumption in single-family houses. *Energy and Buildings,* 29**,** 247-257.
Which. 2011. *History of the UK Energy Market* [Online]. Available: http://www.which.co.uk/switch/energy-advice/history-of-the-energy-market [Accessed 14 June 2011].
Wilcox, R. R. 2005. *Introduction to robust estimation and hypothesis testing,* Burlington, Mass. ; London, Elsevier/Academic Press.
Wilhite, H. & Norgard, J. 2004. Equating efficiency with reduction: a self-deception in energy policy. *Energy and Environment,* 15**,** 991-1009.
Wingfield, J., Bell, M., Miles-Shenton, D., South, T. & Lowe, R. J. 2008. Lessons from Stamford Brook: understanding the gap between designed and real performance. *(Partners in Innovation Project: CI 39/3/663: Evaluating the impact of an enhanced energy performance standard on*

load-bearing masonry domestic construction 8 - Final Report. Leeds, UK: Leeds Metropolitan University.
World Commission on Environment and Development 1987. *Our common future,* Oxford, Oxford University Press.
Xiao, N., Zarnikau, J. & Damien, P. 2007. Testing functional forms in energy modeling: An application of the Bayesian approach to US electricity demand. *Energy Economics,* 29**,** 158-166.
Yao, R. M. & Steemers, K. 2005. A method of formulating energy load profile for domestic buildings in the UK. *Energy and Buildings,* 37**,** 663-671.
Zero Carbon Hub 2011. Allowable solutions for tomorrow's new homes: Towards a workable framework. London: Zero Carbon Hub.
Zimmerman, J. P. 2009. End-use metering campaign in 400 households in Sweden: Assessment of the potential electricity savings. Swedish Energy Agency.

Appendix A: SAS Code

Chapter 5

Code for fuel poverty in households by ONS area classification group (5.5.5.1).

```
/* Import database on estimated numbers of households in England
in fuel poverty by
Lower Layer Super Output Area 2008 */

PROC IMPORT OUT= WORK.fuelpov2008
            DATAFILE=
"C:\Users\steve\Documents\analysis\historicaldata\
fuelpoverty2008\1297-subregional-fuel-poverty-data-2008.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

/* Import ONS Area classifications for each LLSOA */

PROC IMPORT OUT= WORK.areaclass
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\J30A0301_1938_GeoPo
licy_LSOA.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

/* Merge the imported datasets together */

proc sort data=WORK.fuelpov2008;
by LSOA_Code;
run;

proc sort data=WORK.areaclass;
by LSOA_Code;
run;

data work.fuelpovclass;
merge WORK.fuelpov2008 WORK.areaclass;
run;


/* Plot each percent of homes in fuel poverty by area
classification group */

data work.fuelpovclass; set work.fuelpovclass;
label hhfppct ='Percent of homes in fuel poverty 2008';
label group_name ='ONS LLSOA 2001 Area Classification Group';
label supergroup_name ='ONS LLSOA 2001 Area Classification
Supergroup';
run;
```

```
proc sgplot data=work.fuelpovclass;
  hbox hhfppct / category=group_name extreme;

run;
```

Chapter 6

Code for data cleaning to have the base dataset from the 1996 English House Condition Survey where all the cases are assumed to represent households without electric heating (6.2)

```
/* Import of fuel sub-sample of the 1996 English House Condition
Survey with
ONS Area Classification groups attached to some of the cases */

PROC IMPORT OUT= WORK.fuelarea96raw
            DATAFILE=
"C:\Users\steve\Documents\analysis\historicaldata\
EHCS96 public\Energy\Fuel survey\Fuel sample with area
classification R 1996.sav"
            DBMS=SPSS REPLACE;

RUN;

/* Import of meter data from the fuel sub-sample of the 1996
English House Condition Survey */

PROC IMPORT OUT= fuel.Allfuel
            DATAFILE=
"C:\Users\steve\Documents\analysis\historicaldata\
EHCS96 public\Energy\Fuel survey\allfuel.sav"
            DBMS=SPSS REPLACE;

RUN;

/* Import of building data from the fuel sub-sample of the 1996
English House Condition Survey */

PROC IMPORT OUT= fuel.summary
            DATAFILE=
"C:\Users\steve\Documents\analysis\historicaldata\EHCS96
public\Derived
\Combined\summary.sav"
            DBMS=SPSS REPLACE;

RUN;

/* Import of weights from the fuel sub-sample of the 1996 English
House Condition Survey */


PROC IMPORT OUT= fuel.weights
```

```
            DATAFILE=
"C:\Users\steve\Documents\analysis\historicaldata\EHCS96
public\Energy
\Derived\engro96.sav"
            DBMS=SPSS REPLACE;

RUN;

/* Import of household data from the fuel sub-sample of the 1996
English House Condition Survey */

PROC IMPORT OUT= fuel.hhold
            DATAFILE=
"C:\Users\steve\Documents\analysis\historicaldata\EHCS96 public
\Derived\INTSAMP\hhold.sav"
            DBMS=SPSS REPLACE;

RUN;


/* Import of income data from the fuel sub-sample of the 1996
English House Condition Survey */

PROC IMPORT OUT= fuel.netinc96
            DATAFILE=
"C:\Users\steve\Documents\analysis\historicaldata\EHCS96 public
\Derived\INTSAMP\netinc96.sav"
            DBMS=SPSS REPLACE;

RUN;

/* Import of occupancy data from the fuel sub-sample of the 1996
English House Condition Survey */

PROC IMPORT OUT= fuel.occupy
            DATAFILE=
"C:\Users\steve\Documents\analysis\historicaldata\EHCS96 public
\Derived\INTSAMP\occupy.sav"
            DBMS=SPSS REPLACE;

RUN;

/* Import of habitable rroms data from the fuel sub-sample of the
1996 English House Condition Survey */

PROC IMPORT OUT= fuel.rooms
            DATAFILE=
"C:\Users\steve\Documents\analysis\historicaldata\EHCS96 public
\Derived\INTSAMP\rooms96.sav"
            DBMS=SPSS REPLACE;

RUN;

/* Import of heating fuel data from the fuel sub-sample of the
1996 English House Condition Survey */

proc import out=fuel.heat
```

```
datafile="C:\Users\steve\Documents\analysis\historicaldata\EHCS96
public\Energy\Derived\heat96x.sav"
dbms=spss replace; run;


PROC FORMAT LIBRARY=WORK.FORMATS
/* Create SAS database file of formats of categorical data */
CNTLOUT=work.ehcs96formats ;
RUN ;
proc sort data=work.ehcs96formats;
/* Sort by fmtname (Automatic variable created by CNTLOUT process
*/
by fmtname;
run;
/* Export to csv file */
PROC EXPORT DATA= WORK.ehcs96formats
            OUTFILE= "C:\Users\steve\Documents\thesis
temp\ehcs96formats.csv"
DBMS=CSV REPLACE;
PUTNAMES=YES;
RUN;

proc sort data=fuel.heat;
by fieldno;
run;

proc sort data=work.fuelarea96raw;
by fieldno;
run;

proc sort data=fuel.allfuel;
by fieldno;
run;
proc sort data=fuel.summary;
by fieldno;
run;
proc sort data=fuel.weights;
by fieldno;
run;
proc sort data=fuel.netinc96;
by fieldno;
run;
proc sort data=fuel.hhold;
by fieldno;
run;
proc sort data=fuel.occupy;
by fieldno;
run;
proc sort data=fuel.rooms;
by fieldno;
run;

/* Merge all imported data under each case */


data work.ehcs96total;
merge fuel.allfuel fuel.summary fuel.weights fuel.netinc96
fuel.hhold fuel.occupy fuel.rooms fuel.heat fuel.areachar;
```

```
by fieldno;
run;

data work.ehcs96total; set ehcs96total;
if fueldata=. then delete; /* If there is no electricity meter
data, then exclude */
if eannkwh=. then exclude=1; /* If there is no electricity meter
data, then exclude */
if gannkwh=. then exclude=1; /* If there is no gas meter data,
then delete */
if chtyp96x = 6 then exclude=1; /* Reports using electric
floor/ceiling units for heat */
if chtyp96x = 8 then exclude=1; /* Reports using electric storage
for heat */

if chtyp96x = 9 then exclude=1; /* Does not know heating
arrangements */
if chtyp96x = 88 then exclude=1; /* No central heating reported,
likely that there is electric heat used
according to energy fact file data */
if eannkwh=0 then exclude=1; /* If there is no electricity meter
data, then exclude */
rooms_hsize96 = rooms96*hsize96x; /* Interaction term */
if exclude=. then exclude=0;
run;

data ehcs96; set ehcs96total;
if exclude=1 then delete;
run;
```

Code for data cleaning to have the base dataset from the 2008 Living Costs and Food Survey where all the cases are assumed to represent households without electric heating (6.2)

```
/* Import Expenditure figures from 2008 */

PROC IMPORT OUT= fuel.exp2008
            DATAFILE=
"C:\Users\steve\Documents\analysis\historicaldata\LCFS2008spss\sp
ss\
spss12\2008_rawhh_ukanon.sav"
            DBMS=SPSS REPLACE;

/* Import Electricity rates */
proc import out=work.rates2009
datafile="C:\Users\steve\Documents\Papers and thesis\104 - Quant
socioenomic nhood classification\
104 electricity rates 2009.xls"
            DBMS=excel replace;
sheet="sheet1";
getnames=yes; mixed=yes; run;

data work.rates2009_1 ;
```

```
set work.rates2009;
drop f1 f6 f9 f12 f13 f14 f15 f16 f17 f18 f19 f20 f21 f22 f23
f24;
if credit_bill_=0 then delete;
if credit_bill_=. then delete;
rename Credit_Unit_cost=credit;
rename Prepayment_Unit_cost=prepayment;
rename direct_debit_unit_cost=dd;
run;
/* Average out multiple entries per region */
proc sort data=work.rates2009_1;
by region bill_range;
run;

proc means data=work.rates2009_1;
output out=work.rates2009_2 ;
by region bill_range;
run;

data work.rates2009_2; set work.rates2009_2;
if _STAT_ ne 'MEAN' then delete;
drop _N_ _FREQ_ _STAT_ _TYPE_;
run;

/* Transpose to entries by region and payment type */
proc transpose data=work.rates2009_2
out=work.rates2009_3;
by region;
id bill_range;
var credit prepayment dd;
run;

data work.rates2009_3; set work.rates2009_3;
drop _LABEL_;
rename _NAME_=payment_type;
run;

/* Take out all homes with electric heating */

data work.exp2008_1; set fuel.exp2008;
if SerElec > 0 then delete;
keep case year month SampQtr oac gora numadult numchild
numhhldr Nrmsp Nrms2p Nrms3p Nrms4p Nrms5p Nrms6p dvrmsp
elecpay elecrebt ERbtAmt erbtper DVERB EAcAmt EAcAbmt EAcPer
DVEAC EBBSAmt EbbsAbmt EBBSPer DVEBB DSSElecF DSSElecP DSSPer
DVDSSEF DVDSSEP;

run;
```

Code for exclusion of census areas with low levels of central heating (6.2)

```
/* Import of census data on Rooms, Amenities, Central Heating and
Lowest Floor Level (KS19)
for Lower Layer Super Output Areas from the 2001 United Kingdom
Census */
PROC IMPORT OUT= WORK.lsoaheat
```

```
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\KS19
0301_286_GeoPolicy_UK.xls"
            DBMS=EXCEL REPLACE;
     RANGE="LSOAalt$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

/* Establish level to exclude if percent of homes with central
heating > level */
data work.lsoaheat; set work.lsoaheat;
if nocent_bath_p >1 then exclude=99;
if nocent_bath_p >5 then exclude=95;
if nocent_bath_p >10 then exclude=90;
run;
```

Code for Table 1: Cases in the 1996 English House Condition Survey crosstabulated across the 2001 ONS LLSOA Area Classification Supergroup and Government Office Region for England as defined in 1996 (6.2)

```
/* Make table by ONS classification supergroup and government
office region */
ods rtf;
proc freq data=ehcs96;
tables gor_code*supergroup_name / norow nocol;
run; ods rtf close;
```

Code for Table 2: LLSOAs that in the 2001 Census reported more than 95 percent of households having central heating crosstabulated accross 2001 ONS LLSOA Area Classification Supergroup and Government Office Region (6.2)

```
/* Import ONS area classification and merge with 2001 Census data
*/

PROC IMPORT OUT= WORK.allclass
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\J30A
0301_1938_GeoPolicy_LSOA.CSV"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

proc sort data=lsoaheat;
by lsoa_code;
run;

proc sort data=allclass;
```

```
by lsoa_code;
run;

data work.lsoamerge;
merge lsoaheat allclass;
by lsoa_code;
run;

/* Make table by ONS area classification supergroup and
government office region */
ods rtf;
proc freq data=lsoamerge;
where nocent_bath_p<5;
tables gor_code*supergroup_name / norow nocol;
run; ods rtf close;
```

Code for histograms – example given for electricity use in homes without electric heating (6.3)

```
ods graphics on;
ods rtf;
proc univariate data=ehcs96;
var eannkwh;
histogram eannkwh / normal;
(weight GRHFLTRR);
run;
ods graphics off;
ods rtf close;
```

Code for determining number of standard deviations outside of mean electricity use and the interaction term (6.5.1)

```
/* Standardise the dependent variable, electricity, to a mean of
zero and a standard deviation of 1 */
data ehcs96z; set ehcs96;
zeannkwh = eannkwh;
zrooms_hsize96 = rooms_hsize96;
label eannkwh = "Annual non-heating end-use energy (kWh of
electricity, 1996)";
label rooms_hsize96 = "Interaction term of the number of rooms
and number of occupants (1996)";
label zeannkwh = "Z-score of non-heating end-use energy";
label zrooms_hsize96 = "Z-score of interaction term";
run;

proc standard data=ehcs96z
mean=0 std=1
out=ehcsoutliers;
var zeannkwh zrooms_hsize96;
run;

/* Label numbers of standard deviations beyond the mean */

data ehcsoutliers; set ehcsoutliers;
```

```
if abs(zeannkwh) <= 1.96 then outliere = 0;
if abs(zeannkwh) > 1.96 then outliere = 1;
if abs(zeannkwh) > 2.58 then outliere = 2;
if abs(zeannkwh) > 3.59 then outliere = 3;
if zeannkwh =. then outliere = -1;
if zeannkwh =0 then outliere = -1;
if abs(zrooms_hsize96) <= 1.96 then outliers = 0;
if abs(zrooms_hsize96) > 1.96 then outliers = 1;
if abs(zrooms_hsize96) > 2.58 then outliers = 2;
if abs(zrooms_hsize96) > 3.59 then outliers = 3;
if zrooms_hsize96 =. then outliers = -1;
if zrooms_hsize96 =0 then outliers = -1;
label outliere = "Number of whole standard deviations outside of
the mean energy use";
label outliers = "Number of whole standard deviations outside of
the mean interaction term";
if outliere>1 then outlierall=1;
if outliere=-1 then outlierall=1;
if outliers>1 then outlierall=1;
if outliers=-1 then outlierall=1;
if outlierall=. then outlierall=0;
label outlierall = "Excluded cases";
rpp = rooms96/hsize96x;
if hsize96x>rooms96 then over=1; else over=0;
if (4*hsize96x)<rooms96 then under=1; else under=0;
sqrt_eannkwh = sqrt(eannkwh);
keep sqrt_eannkwh eannkwh rooms_hsize96 zeannkwh zrooms_hsize96
outliere outliers outlierall rooms96 hsize96x rpp over under;
run;
```

Code for Box-Cox analysis (6.5.2)

```
ods graphics on;

   title2 'Default Options';

   proc transreg data=ehcs96 test;
      model BoxCox(eannkwh) = identity(rooms_hsize96);
   run;
   ods graphics off;

   ods graphics on;

   title2 'Default Options';

   proc transreg data=outlier2 test;
      model BoxCox(eannkwh) = identity(rooms_hsize96);
   run;
   ods graphics off;

   ods graphics on;

   title2 'Default Options';

   proc transreg data=outlier2 test;
      model BoxCox(eannkwh) = log(eannkwh);
   run;
```

```
ods graphics off;
```

Code for tests of skewness and kurtosis, histograms, quantile-quantile plots using the normtrans macro (6.5.2)

```
%normtrans (ehcs96, eannkwh, rooms_hsize96, GRHFLTRR);
/* Dataset from the fuel sub-sample EHCS with Annual electricity
use, interaction term, and weight from the fuel subsample of the
EHCS */


%MACRO normtrans (data, var, vartwo, weight);

data data2; set &data;
z&var = &var;
z&vartwo = &vartwo;
run;

proc standard data=data2
mean=0 std=1
out=outlierset;
var z&var z&vartwo;
run;

data outlierset; set outlierset;
if abs(z&var) <= 1.96 then outlier = 0;
if abs(z&var) > 1.96 then outlier = 1;
if abs(z&var) > 2.58 then outlier = 2;
if abs(z&var) > 3.59 then outlier = 3;
if z&var =. then outlier = -1;
if z&var =0 then outlier = -1;
if abs(z&vartwo) <= 1.96 then leverage = 0;
if abs(z&vartwo) > 1.96 then leverage = 1;
if abs(z&vartwo) > 2.58 then leverage = 2;
if abs(z&vartwo) > 3.59 then leverage = 3;
if z&vartwo =. then leverage = -1;
if z&vartwo =0 then leverage = -1;
if outlier>1 then toexclude=1;
if outlier=-1 then toexclude=1;
if leverage>1 then toexclude=1;
if leverage=-1 then toexclude=1;
if leverage=. then toexclude=0;
keep &var &vartwo &weight z&var z&vartwo outlier leverage
toexclude;
run;

proc sort data=outlierset;
by toexclude;
run;

/* Eliminate outliers above 2.58 */

data outlier2; set outlierset;
if toexclude=1 then delete;
```

```
run;
/* Normality test program */

ods rtf;
ods graphics on;
proc univariate data=data2;
var &var;
histogram &var / normal;
qqplot &var;
ppplot &var /normal;
output out=transNOoutliersIN
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
ods graphics off;
data transNOoutliersIN; set transNOoutliersIN;
format name $30.; name = "transNOoutliersIN";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;



/* Positive skew - Log and Sqrt and 4th root t_formations */
data logdata; set data2;
if &var=. then delete;
if &var=0 then delete;
logvar = log(&var);
run;

data sqrtdata; set data2;
if &var=. then delete;
if &var=0 then delete;
sqrtvar = sqrt(&var);
run;

data fthrtdata; set data2;
if &var=. then delete;
if &var=0 then delete;
fthrtvar = (&var)**(1/4);
run;

data sqdata; set data2;
if &var=. then delete;
if &var=0 then delete;
sqvar = (&var)**2;
run;

ods graphics on;
proc univariate data=logdata;
var logvar;
```

```
histogram logvar / normal;
qqplot logvar;
ppplot logvar /normal;
output out=transLOGoutliersIN
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
ods graphics off;
data transLOGoutliersIN; set transLOGoutliersIN;
format name $20.;
format name $30.; name = "transLOGoutliersIN";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

ods graphics on;
proc univariate data=sqrtdata;
var sqrtvar;
histogram sqrtvar / normal;
qqplot sqrtvar;
ppplot sqrtvar /normal;
output out=transSQRToutliersIN
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
ods graphics off;
data transSQRToutliersIN; set transSQRToutliersIN;
format name $30.; name = "transSQRToutliersIN";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

ods graphics on;
proc univariate data=fthrtdata;
var fthrtvar;
histogram fthrtvar / normal;
qqplot fthrtvar;
ppplot fthrtvar /normal;
output out=transFTHRToutliersIN
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
ods graphics off;
data transFTHRToutliersIN; set transFTHRToutliersIN;
```

```
format name $30.; name = "transFTHRToutliersIN";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;


ods graphics on;
proc univariate data=sqdata;
var sqvar;
histogram sqvar / normal;
qqplot sqvar;
ppplot sqvar /normal;
output out=transSQoutliersIN
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
ods graphics off;
data transSQoutliersIN; set transSQoutliersIN;
format name $30.; name = "transSQoutliersIN";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

ods graphics on;
proc univariate data=outlier2;
var &var;
histogram &var / normal;
qqplot &var;
ppplot &var /normal;
output out=transNOoutliersOUT
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
ods graphics off;
data transNOoutliersOUT; set transNOoutliersOUT;
format name $30.; name = "transNOoutliersOUT";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;


data outlog; set outlier2;
```

```
if &var=. then delete;
if &var=0 then delete;
logvar = log(&var);
run;

data outsqrt; set outlier2;
if &var=. then delete;
if &var=0 then delete;
sqrtvar = sqrt(&var);
run;

data outsq; set outlier2;
if &var=. then delete;
if &var=0 then delete;
sqvar = (&var)**2;
run;

data outfthrt; set outlier2;
if &var=. then delete;
if &var=0 then delete;
fthrtvar = (&var)**(1/4);
run;

ods graphics on;
proc univariate data=outlog;
var logvar;
histogram logvar / normal;
qqplot logvar;
ppplot logvar /normal;
output out=transLOGoutliersOUT
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
ods graphics off;
data transLOGoutliersOUT; set transLOGoutliersOUT;
format name $30.; name = "transLOGoutliersOUT";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

ods graphics on;
proc univariate data=outsqrt;
var sqrtvar;
histogram sqrtvar / normal;
qqplot sqrtvar;
ppplot sqrtvar /normal;
output out=transSQRToutliersOUT
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
```

```
ods graphics off;
data transSQRToutliersOUT; set transSQRToutliersOUT;
format name $30.; name = "transSQRToutliersOUT";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

ods graphics on;
proc univariate data=outfthrt;
var fthrtvar;
histogram fthrtvar / normal;
qqplot fthrtvar;
ppplot fthrtvar /normal;
output out=transFTHRToutliersOUT
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
ods graphics off;
data transFTHRToutliersOUT; set transFTHRToutliersOUT;
format name $30.; name = "transFTHRToutliersOUT";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

ods graphics on;
proc univariate data=outsq;
var sqvar;
histogram sqvar / normal;
qqplot sqvar;
ppplot sqvar /normal;
output out=transSQoutliersOUT
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
ods graphics off;
data transSQoutliersOUT; set transSQoutliersOUT;
format name $30.; name = "transSQoutliersOUT";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

/* Add weights */
```

```
proc univariate data=data2;
var &var; weight &weight;
output out=transNOoutliersIN_w
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
data transNOoutliersIN_w; set transNOoutliersIN_w;
format name $30.; name = "transNOoutliersIN_w";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

proc univariate data=logdata;
var logvar; weight &weight;
output out=transLOGoutliersIN_w
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
data transLOGoutliersIN_w; set transLOGoutliersIN_w;
format name $30.; name = "transLOGoutliersIN_w";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

proc univariate data=sqrtdata;
var sqrtvar; weight &weight;
output out=transSQRToutliersIN_w
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
data transSQRToutliersIN_w; set transSQRToutliersIN_w;
format name $30.; name = "transSQRToutliersIN_w";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

proc univariate data=fthrtdata;
var fthrtvar; weight &weight;
output out=transFTHRToutliersIN_w
n=Number
```

```
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
data transFTHRToutliersIN_w; set transFTHRToutliersIN_w;
format name $30.; name = "transFTHRToutliersIN_w";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

proc univariate data=outlier2;
var &var; weight &weight;
output out=transNOoutliersOUT_w
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
data transNOoutliersOUT_w; set transNOoutliersOUT_w;
format name $30.; name = "transNOoutliersOUT_w";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

proc univariate data=outlog;
var logvar; weight &weight;
output out=transLOGoutliersOUT_w
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
data transLOGoutliersOUT_w; set transLOGoutliersOUT_w;
format name $30.; name = "transLOGoutliersOUT_w";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

proc univariate data=outsqrt;
var sqrtvar; weight &weight;
output out=transSQRToutliersOUT_w
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
```

```
data transSQRToutliersOUT_w; set transSQRToutliersOUT_w;
format name $30.; name = "transSQRToutliersOUT_w";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;

proc univariate data=outfthrt;
var fthrtvar; weight &weight;
output out=transFTHRToutliersOUT_w
n=Number
mean=Mean
skewness=Skewness
kurtosis=Kurtosis;
run;
data transFTHRToutliersOUT_w; set transFTHRToutliersOUT_w;
format name $30.; name = "transFTHRToutliersOUT_w";
seskew=((((6*Number)*(Number-1))/((Number-
2)*(Number+1)*(Number+3)))**0.5);
sekurt=(2*seskew*((Number**2*(2-1))/((Number-
3)*(Number+5)))**0.5);
zkurt=Kurtosis/sekurt;
zskew=Skewness/seskew;
run;


/* Output table of skewness and kurtosis */

data normality;
set transNOoutliersIN
transLOGoutliersIN
transSQRToutliersIN
transFTHRToutliersIN
transNOoutliersOUT
transLOGoutliersOUT
transSQRToutliersOUT
transFTHRToutliersOUT
transNOoutliersIN_w
transLOGoutliersIN_w
transSQRToutliersIN_w
transFTHRToutliersIN_w
transNOoutliersOUT_w
transLOGoutliersOUT_w
transSQRToutliersOUT_w
transFTHRToutliersOUT_w;
run;

proc print data=normality; run;

ods rtf close;

 %MEND normtrans;
```

Code for conversion of electricity values in the 1996 EHCS fuel sub-sample to 2008 values using 2008 Living Costs and Food Survey (6.6)

```
/* Estimate electricity use from each bill */
data exp2008_merge2; set exp2008_merge;
if EAcPer=5 and payment_type='credit' then elec =
(EAcAmt+ERbtAmt) / average * 100 ;
if EAcPer=13 and payment_type='credit' then elec =
(EAcAmt+ERbtAmt) / average * 100 / 3 ;
if EAcPer=5 and payment_type='dd' then elec = (EAcAmt+ERbtAmt) /
average * 100 ;
if EAcPer=13 and payment_type='dd' then elec = (EAcAmt+ERbtAmt) /
average * 100 / 3 ;
if EBBSPer=5 and payment_type='credit' then elec = (EBBSAmt) /
average * 100 ;
if EBBSPer=13 and payment_type='credit' then elec = (EBBSAmt) /
average * 100 / 3 ;
if elec=. then delete;
array expmonth {12} expmonth1-expmonth12;
format expmonth1-expmonth12 8.2;
do i=1 to 12;
if month=i then expmonth{i} = elec;
end;
run;

proc sort data=work.ehcs96;
by efirdate;
run;
/*Elimate outliers */

%inc "C:\Users\steve\Documents\thesis temp\Chapter 6 -
Interpretation\outliermacro.sas";
%outlier (ehcs96, eannkwh, rooms_hsize96);

%MACRO outlier (data, var, vartwo);
data data2; set &data;
z&var = &var;
z&vartwo = &vartwo;
run;

proc standard data=data2
mean=0 std=1
out=outlierset;
var z&var z&vartwo;
run;

data outlierset; set outlierset;
if abs(z&var) <= 1.96 then outlier = 0;
if abs(z&var) > 1.96 then outlier = 1;
if abs(z&var) > 2.58 then outlier = 2;
if abs(z&var) > 3.59 then outlier = 3;
if z&var =. then outlier = -1;
if z&var =0 then outlier = -1;
if abs(z&vartwo) <= 1.96 then leverage = 0;
if abs(z&vartwo) > 1.96 then leverage = 1;
if abs(z&vartwo) > 2.58 then leverage = 2;
if abs(z&vartwo) > 3.59 then leverage = 3;
if z&vartwo =. then leverage = -1;
if z&vartwo =0 then leverage = -1;
if outlier>1 then toexclude=1;
```

```
if outlier=-1 then toexclude=1;
if leverage>1 then toexclude=1;
if leverage=-1 then toexclude=1;
if leverage=. then toexclude=0;
run;


proc sort data=outlierset;
by toexclude;
run;

/* Eliminate outliers above 2.58 */

data outlier2; set outlierset;
if toexclude=1 then delete;
run;

%Mend outlier;


/* Assign beginning month of quarter to all quarters */
data work.ehcs96_dates; set work.outlier2;
if elecpres ne 1 then delete;
array ekwh {9} ekwh1-ekwh9;
array qmonth {12} qmonth1-qmonth12;
array qstart {39} qstart1-qstart39;
array ekwhmonth {12} ekwhmonth1-ekwhmonth12;
format qstart1-qstart39 8.2 ;
format ekwhmonth1-ekwhmonth12 8.2;
do i = 1 to 12;
qmonth{i} = month(efirdate)+11+i;
end;
do i=1 to 9; /*Replace zeros with missing values*/
if ekwh{i}=0 then ekwh{i}=.;
end;
/* 1st year */
do i = 13 to 24;
if qmonth1=i then
do; qstart{i}=mean(ekwh1,ekwh5,ekwh9)/3;
qstart{i-1}=mean(ekwh1,ekwh5,ekwh9)/3;
qstart{i-2}=mean(ekwh1,ekwh5,ekwh9)/3; end;
if qmonth4=i then
do; qstart{i}=mean(ekwh2,ekwh6)/3;
qstart{i-1}=mean(ekwh2,ekwh6)/3;
qstart{i-2}=mean(ekwh2,ekwh6)/3; end;
if qmonth7=i then
do; qstart{i}=mean(ekwh3,ekwh7)/3;
qstart{i-1}=mean(ekwh3,ekwh7)/3;
qstart{i-2}=mean(ekwh3,ekwh7)/3; end;
if qmonth10=i then
do; qstart{i}=mean(ekwh4,ekwh8)/3;
qstart{i-1}=mean(ekwh4,ekwh8)/3;
qstart{i-2}=mean(ekwh4,ekwh8)/3; end;
end;
/* 2nd year */
do i = 25 to 36;
if qmonth1=i then
do; qstart{i}=mean(ekwh1,ekwh5,ekwh9)/3;
```

```
qstart{i-1}=mean(ekwh1,ekwh5,ekwh9)/3;
qstart{i-2}=mean(ekwh1,ekwh5,ekwh9)/3; end;
if qmonth4=i then
do; qstart{i}=mean(ekwh2,ekwh6)/3;
qstart{i-1}=mean(ekwh2,ekwh6)/3;
qstart{i-2}=mean(ekwh2,ekwh6)/3; end;
if qmonth7=i then
do; qstart{i}=mean(ekwh3,ekwh7)/3;
qstart{i-1}=mean(ekwh3,ekwh7)/3;
qstart{i-2}=mean(ekwh3,ekwh7)/3; end;
if qmonth10=i then
do; qstart{i}=mean(ekwh4,ekwh8)/3;
qstart{i-1}=mean(ekwh4,ekwh8)/3;
qstart{i-2}=mean(ekwh4,ekwh8)/3; end;
end;
do i=1 to 12;
ekwhmonth{i}=mean(qstart{i}, qstart{i+12}, qstart{i+24});
if ekwhmonth{i}<0 then delete;
end;
drop qstart1-qstart39 qmonth1-qmonth12 i;
run;

proc means data=exp2008_merge2;
output out=expmeans;
var expmonth1-expmonth12;
run;

proc transpose data=expmeans
out=expmeanst; run;

data expmeans; set expmeans;
if _STAT_ ne "MEAN" then delete;
drop _TYPE_ _FREQ_ _STAT_;
run;


data expmeanst; set expmeanst;
if _NAME_ = "_TYPE_" then delete;
if _NAME_ = "_FREQ_" then delete;
keep col5; run;

proc transpose data=expmeanst out=monthstd; run;

data monthstd; set monthstd;
array col {12} col1-col12;
array stdmonth {12} stdmonth1-stdmonth12;
do i=1 to 12;
stdmonth{i}=col{i}; end;
keep stdmonth1-stdmonth12;
run;

data echs96_08means;
if _N_ = 1 then set expmeans;
if _N_ = 1 then set monthstd ;
set ehcs96_dates;
run;

proc standard data=echs96_08means out=ehcs96_std noprint
```

```
mean=0 std=1;
var ekwhmonth1-ekwhmonth12;
run;

data echs96_08conv; set ehcs96_std;
array ekwhmonthconv {12} ekwhmonthconv1-ekwhmonthconv12;
array expmonth {12} expmonth1-expmonth12;
array ekwhmonth {12} ekwhmonth1-ekwhmonth12;
array stdmonth {12} stdmonth1-stdmonth12;
format ekwhmonthconv1-ekwhmonthconv12 8.2;
do i=1 to 12;
ekwhmonthconv{i}=expmonth{i}+ekwhmonth{i}*stdmonth{i};
end;
eannkwh_2008 = sum(of ekwhmonthconv1-ekwhmonthconv12);
run;

/* Export tables of average electricity use by month in 1996 and
2008 */

ods rtf;
proc means data=ehcs96_dates mean median;
var ekwhmonth1-ekwhmonth12;
run;

proc means data=echs96_08conv mean median;
var ekwhmonthconv1-ekwhmonthconv12;
run;
ods rtf close;
```

Chapter 7

Code for the running of the single-level model. Example using the annual option (7.2)

```
/* Using the adjusted data for 2008
Standardise the dependent variable, electricity, to a mean of
zero and a standard deviation of 1 */
data ehcs96z; set echs96_08conv;
sqrt_eannkwh = sqrt(eannkwh);
zeannkwh = sqrt_eannkwh;
zrooms_hsize96 = rooms_hsize96;
label eannkwh = "Annual non-heating end-use energy (kWh of
electricity, 1996)";
label rooms_hsize96 = "Interaction term of the number of rooms
and number of occupants (1996)";
label zeannkwh = "Z-score of square-root transformed non-heating
end-use energy";
label zrooms_hsize96 = "Z-score of interaction term";
run;

proc standard data=ehcs96z
mean=0 std=1
out=ehcsoutliers;
var zeannkwh zrooms_hsize96;
run;
```

```
data ehcsoutliers; set ehcsoutliers;
if abs(zeannkwh) <= 1.96 then outliere = 0;
if abs(zeannkwh) > 1.96 then outliere = 1;
if abs(zeannkwh) > 2.58 then outliere = 2;
if abs(zeannkwh) > 3.59 then outliere = 3;
if zeannkwh =. then outliere = -1;
if zeannkwh =0 then outliere = -1;
if abs(zrooms_hsize96) <= 1.96 then outliers = 0;
if abs(zrooms_hsize96) > 1.96 then outliers = 1;
if abs(zrooms_hsize96) > 2.58 then outliers = 2;
if abs(zrooms_hsize96) > 3.59 then outliers = 3;
if zrooms_hsize96 =. then outliers = -1;
if zrooms_hsize96 =0 then outliers = -1;
label outliere = "Number of whole standard deviations outside of
the mean energy use";
label outliers = "Number of whole standard deviations outside of
the mean interaction term";
if outliere>1 then outlierall=1;
if outliere=-1 then outlierall=1;
if outliers>1 then outlierall=1;
if outliers=-1 then outlierall=1;
if outlierall=. then outlierall=0;
label outlierall = "Excluded cases";
rpp = rooms96/hsize96x;
if hsize96x>rooms96 then over=1; else over=0;
if (4*hsize96x)<rooms96 then under=1; else under=0;
keep sqrt_eannkwh eannkwh rooms_hsize96 zeannkwh zrooms_hsize96
outliere outliers outlierall rooms96 hsize96x rpp over under;
run;

/* Exclude outliers more than 2 standard deviations away from the
mean */

data ehcsoutliers_ex; set ehcsoutliers;
if outlierall = 1 then delete;
/*rooms_hsize2 = rooms_hsize96**2*/;
run;

/* Ordinary least squares regression */

ods graphics on;
ods rtf;
proc reg data=ehcsoutliers_ex corr simple;
model sqrt_eannkwh = rooms_hsize96
/* sqrt_eannkwh is square root transformed
rooms_hsize is the interaction term */;
output out = outputfile
residual = outputresidual
predicted = outputpredicted
student = outputstudent
rstudent = outputrstudent
dffits = outputdffits
cookd = outputcooksd
h = outputleverage;
run; ods rtf close; quit;

/* Taking out residuals more than 2 s.d. */
```

```
data studentout; set outputfile;
if outputstudent > 2 then delete;
if outputstudent < -2 then delete;
/* rooms_hsize2 = rooms_hsize96**2; */
run;

/* Second run of OLS regression */

ods graphics on;
ods rtf;
proc reg data=studentout corr simple;
model sqrt_eannkwh = rooms_hsize96
/* sqrt_eannkwh is square root transformed
rooms_hsize is the interaction term */;
output out = outputfile2
residual = outputresidual
predicted = outputpredicted
student = outputstudent
rstudent = outputrstudent
dffits = outputdffits
cookd = outputcooksd
h = outputleverage;
run; ods rtf close; quit;
```

Code for validation of single-level model using annual model set at 2008 (7.3)

```
/* Combine reference lists of lower layer super output area to
local authority */
PROC IMPORT OUT= WORK.oaref1
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\OA_L
SOA_MSOA_LA_Apr05_part1.xls"
            DBMS=EXCEL REPLACE;
     RANGE="OA_LSOA_MSOA_LA_Apr05_part1$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

PROC IMPORT OUT= WORK.oaref2
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\OA_L
SOA_MSOA_LA_Apr05_part2.xls"
            DBMS=EXCEL REPLACE;
     RANGE="OA_LSOA_MSOA_LA_Apr05_part2$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

PROC IMPORT OUT= WORK.oaref3
```

```
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\OA_L
SOA_MSOA_LA_Apr05_part3.xls"
            DBMS=EXCEL REPLACE;
     RANGE="OA_LSOA_MSOA_LA_Apr05_part3$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

data oaref; set oaref1 oaref2 oaref3;
run;

/* Import crosstabs */
PROC IMPORT OUT= WORK.OAcrosstab
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\oa_c
rosstab.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

PROC IMPORT OUT= WORK.uacrosstab
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\la_c
rosstab.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;


PROC IMPORT OUT= WORK.districtcrosstab
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\district_c
rosstab.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

data lacrosstab; set districtcrosstab uacrosstab; run;

data oacrosstab; set oacrosstab;
rename zone_code = oa_code;
run;

/* Create a crosstab by output area then total up to LA to check
zeros */
data oacrosstab_2;
merge oacrosstab oaref;
by oa_code;
run;

proc means data=oacrosstab_2 noprint;
```

```
output out = oacrosstab_3
sum(cs0510001-cs0510150)=cs_sum0510001-cs_sum0510150;
by la_code; run;

proc sort data=oacrosstab_2; by lsoa_code; run;

/* Sum all of the output areas into lower layer super output
areas */
proc means data=oacrosstab_2 noprint;
output out = oacrosstab_lsoa
sum(cs0510001-cs0510150)=cs_sum0510001-cs_sum0510150;
by lsoa_code; run;

/* Take in estimated number of households 2008 from fuel pov and
distribute */

PROC IMPORT OUT= WORK.households2008
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\1297-subregional-
fuel-poverty-data-2008.xls"
            DBMS=EXCEL REPLACE;
     RANGE="LSOA$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
/* Transfer to estimated number of households in 2008 */
proc sort data=households2008; by lsoa_code; run;
data oacrosstab_lsoa_2008;
merge oacrosstab_lsoa households2008; by lsoa_code; run;

data oacrosstab_lsoa_2008; set oacrosstab_lsoa_2008;
array housingbefore {149} cs_sum0510002-cs_sum0510150;
array housingafter {149} cs_new0510002-cs_new0510150;
do i=1 to 149;
housingafter {i} = housingbefore {i} * hh2008 / cs_sum0510001;
end;
run;

data oacrosstab_lsoa_2008; set oacrosstab_lsoa_2008;
drop f12;
electotal_2001 =
cs_sum0510008*(1.1*1**2*1**2 + 89.7*1*1+1820)+
cs_sum0510009*(1.1*1**2*2**2 + 89.7*1*2+1820)+
cs_sum0510010*(1.1*1**2*3.73**2 + 89.7*1*3.73+1820)+
cs_sum0510011*(1.1*1**2*5.52**2 + 89.7*1*5.52+1820)+
cs_sum0510012*(1.1*1**2*7.30**2 + 89.7*1*7.30+1820)+
cs_sum0510014*(1.1*2**2*1**2 + 89.7*2*1+1820)+
cs_sum0510015*(1.1*2**2*2**2 + 89.7*2*2+1820)+
cs_sum0510016*(1.1*2**2*3.73**2 + 89.7*2*3.73+1820)+
cs_sum0510017*(1.1*2**2*5.52**2 + 89.7*2*5.52+1820)+
cs_sum0510018*(1.1*2**2*7.30**2 + 89.7*2*7.30+1820)+
cs_sum0510020*(1.1*3.5**2*1**2 + 89.7*3.5*1+1820)+
cs_sum0510021*(1.1*3.5**2*2**2 + 89.7*3.5*2+1820)+
cs_sum0510022*(1.1*3.5**2*3.73**2 + 89.7*3.5*3.73+1820)+
cs_sum0510023*(1.1*3.5**2*5.52**2 + 89.7*3.5*5.52+1820)+
```

```
cs_sum0510024*(1.1*3.5**2*7.30**2 + 89.7*3.5*7.30+1820)+
cs_sum0510026*(1.1*5.52**2*1**2 + 89.7*5.52*1+1820)+
cs_sum0510027*(1.1*5.52**2*2**2 + 89.7*5.52*2+1820)+
cs_sum0510028*(1.1*5.52**2*3.73**2 + 89.7*5.52*3.73+1820)+
cs_sum0510029*(1.1*5.52**2*5.52**2 + 89.7*5.52*5.52+1820)+
cs_sum0510030*(1.1*5.52**2*7.30**2 + 89.7*5.52*7.30+1820);

electotal_2008 =
cs_new0510008*(1.12*1**2*1**2 + 90.5*1*1+1820)+
cs_new0510009*(1.12*1**2*2**2 + 90.5*1*2+1820)+
cs_new0510010*(1.12*1**2*3.73**2 + 90.5*1*3.73+1820)+
cs_new0510011*(1.12*1**2*5.52**2 + 90.5*1*5.52+1820)+
cs_new0510012*(1.12*1**2*7.30**2 + 90.5*1*7.30+1820)+
cs_new0510014*(1.12*2**2*1**2 + 90.5*2*1+1820)+
cs_new0510015*(1.12*2**2*2**2 + 90.5*2*2+1820)+
cs_new0510016*(1.12*2**2*3.73**2 + 90.5*2*3.73+1820)+
cs_new0510017*(1.12*2**2*5.52**2 + 90.5*2*5.52+1820)+
cs_new0510018*(1.12*2**2*7.30**2 + 90.5*2*7.30+1820)+
cs_new0510020*(1.12*3.5**2*1**2 + 90.5*3.5*1+1820)+
cs_new0510021*(1.12*3.5**2*2**2 + 90.5*3.5*2+1820)+
cs_new0510022*(1.12*3.5**2*3.73**2 + 90.5*3.5*3.73+1820)+
cs_new0510023*(1.12*3.5**2*5.52**2 + 90.5*3.5*5.52+1820)+
cs_new0510024*(1.12*3.5**2*7.30**2 + 90.5*3.5*7.30+1820)+
cs_new0510026*(1.12*5.52**2*1**2 + 90.5*5.52*1+1820)+
cs_new0510027*(1.12*5.52**2*2**2 + 90.5*5.52*2+1820)+
cs_new0510028*(1.12*5.52**2*3.73**2 + 90.5*5.52*3.73+1820)+
cs_new0510029*(1.12*5.52**2*5.52**2 + 90.5*5.52*5.52+1820)+
cs_new0510030*(1.12*5.52**2*7.30**2 + 90.5*5.52*7.30+1820);

run;



/* Compare with the reliable output areas */

PROC IMPORT OUT= WORK.elec2008
            DATAFILE=
"C:\Users\steve\Documents\analysis\nhoodstats\LLSOA_Master_2008_a
lt.xls"
            DBMS=EXCEL REPLACE;
     RANGE="DOM_ELEC$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

proc sort data=elec2008; by lsoa_code; run;

data crosstablinear;
merge oacrosstab_lsoa_2008 lsoamerge elec2008;
by lsoa_code;
run;

data crosstablinear; set crosstablinear;
if exclude>90 then delete;
if cs_sum0510001 = . then delete;
```

```
diff = electotal_2008 - con_elec_ord;
diffpct = diff/con_elec_ord;
run;

proc sort data=crosstablinear; by supergroup_name; run;

proc means data=crosstablinear;
output out=newsum
sum(electotal_2008)=estsum
sum(con_elec_ord)=elecsum
sum(diff)=diffsum
;
by supergroup_name;
run;


data newsum; set newsum;
pct = diffsum/elecsum;
label supergroup_name = '2001 Area Classification Supergroup
Name';
label estsum = 'Model estimate of electricity use in 2008 (kWh)';
label _FREQ_ = 'Number of LSOAs with central heating > 95%;';
label elecsum = 'Actual electricity use recorded in 2008 (kWh)';
label diffsum = 'Difference between estimate and actual use
(kWh)';
label pct = 'Difference between estimate and actual use
(percent)';
run;

/* Output means by supergroup */
proc means data=crosstablinear;
by supergroup_name;
run;
```

Code for the multilevel model (7.4)

```
/* Add in means of interaction term from groups */
/* This dataset is the mean people per household and rooms per
household from the 2001 United Kingdom Census */
proc import out=work.meansize
datafile="C:\Users\steve\Documents\analysis\nhoodstats\KS190301_2
86_GeoPolicy_UK.xls"
            DBMS=excel replace;
sheet="LSOAalt";
getnames=yes; mixed=yes; run;

/* This is the dataset of LLSOA Area Classifications */
proc import out=work.class_lsoa
datafile="C:\Users\steve\Documents\analysis\nhoodstats\J30A0301_1
938_GeoPolicy_LSOA.csv"
dbms=csv replace;
run;

/* More mean per household in each LLSOA with its Area
Classification */
proc sort data=meansize; by lsoa_code; run;
proc sort data=class_lsoa; by lsoa_code; run;
```

```
data meansize_class;
merge meansize class_lsoa;
by lsoa_code;
rp_mean = pph*rph;
if gor_code=" " then delete;
run;

/* Create mean and median values for interaction term for
supergroups
and then for groups*/
proc sort data=meansize_class; by supergroup_code; run;

proc means data=meansize_class ;
output out=meansize_class_super
mean=super_mean_raw median=super_median_raw
;
var rp_mean; by supergroup_code;
run;

proc sort data=meansize_class; by group_code; run;

proc means data=meansize_class;
output out=meansize_class_group
mean=group_mean_raw median=group_median_raw;
var rp_mean; by group_code;
run;

/* Centre studentout, meansize_class_super, and
meansize_class_group */


data c_studentout; set studentout;
c_rooms_hsize96 = rooms_hsize96 - 12.58;
if urban96x=1 then urban=1;
if urban96x=2 then urban=0;
if area96x=3 then suburb=1; else suburb=0;
if house96x=1 then house=1; else house=0;
if hage296x=1 then elderly=1; else elderly=0;
if latyp96x=1 then prewar=1; else prewar=0;
run;

data c_meansize_class_super; set meansize_class_super;
super_mean = super_mean_raw - 12.58;
run;

data c_meansize_class_group; set meansize_class_group;
group_mean = group_mean_raw - 12.58;
run;


/* Add in the reduced dataset from previous - 1996 decennial
model*/

proc sort data=c_studentout; by supergroup_code; run;

data withmeans1;
merge c_studentout c_meansize_class_super; by supergroup_code;
```

```
run;

proc sort data=c_studentout; by group_code; run;
proc sort data=withmeans1; by group_code; run;

data withmeans2;
merge withmeans1 c_meansize_class_group; by group_code;
run;

/* Try 2008 annual model */
proc sort data=echs96_08conv; by supergroup_code; run;

data withmeans08_1;
merge echs96_08conv meansize_class_super; by supergroup_code;
run;

proc sort data=echs96_08conv; by supergroup_code; run;
proc sort data=withmeans08_1; by supergroup_code; run;

data withmeans08_2;
merge withmeans08_1 meansize_class_super; by supergroup_code;
run;

/* Initial model - analysis of variance */



ods rtf file="C:\Users\steve\Documents\Papers and thesis\Chapter
7 - Critical assessment\Ch7 multilevel centred supergroup.rtf";
/* Unconditional means model */
proc mixed data = withmeans2 covtest noclprint;
   class supergroup_code;
   model sqrt_eannkwh = / solution;
   random intercept / subject = supergroup_code;
run;


/* Taking into account level2 means only - ecological model */
proc mixed data = withmeans2 covtest noclprint;
   class supergroup_code;
   model sqrt_eannkwh = super_mean / solution ddfm = bw;
   random intercept / subject = supergroup_code;
run;


data withmeans96; set withmeans2;
hsize_s = c_rooms_hsize96 - super_mean;
run;
/* Include effect of individual-level predictor of household size
*/
proc mixed data = withmeans96 noclprint covtest;
  class supergroup_code;
  model sqrt_eannkwh = hsize_s / solution ddfm = bw notest;
  random intercept hsize_s / subject = supergroup_code type = un
gcorr;
run;
```

```
/* Include both effects
proc mixed data = withmeans96 noclprint covtest noitprint;
  class supergroup_code;
  model sqrt_eannkwh = super_mean hsize_s hsize_s*super_mean /
solution ddfm = bw notest;
  random intercept hsize_s / subject = supergroup_code type = un
;
run; */

/* Test binomial predictors:
Surburban / not suburban
Urban / not urban
House / not house (flat)
Head of household over 60 / not over 60
House built before 1946 / after 1945 */


proc mixed data = withmeans96 noclprint covtest noitprint;
  class supergroup_code;
  model sqrt_eannkwh = super_mean suburb hsize_s
hsize_s*super_mean hsize_s*suburb  / solution ddfm = bw notest;
  random intercept hsize_s / subject = supergroup_code type = un
;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
  class supergroup_code;
  model sqrt_eannkwh = super_mean urban hsize_s
hsize_s*super_mean hsize_s*urban  / solution ddfm = bw notest;
  random intercept hsize_s / subject = supergroup_code type = un
;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
  class supergroup_code;
  model sqrt_eannkwh = super_mean house hsize_s
hsize_s*super_mean hsize_s*house  / solution ddfm = bw notest;
  random intercept hsize_s / subject = supergroup_code type = un
;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
  class supergroup_code;
  model sqrt_eannkwh = super_mean elderly hsize_s
hsize_s*super_mean hsize_s*elderly  / solution ddfm = bw notest;
  random intercept hsize_s / subject = supergroup_code type = un
;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
  class supergroup_code;
  model sqrt_eannkwh = super_mean prewar hsize_s
hsize_s*super_mean hsize_s*prewar  / solution ddfm = bw notest;
  random intercept hsize_s / subject = supergroup_code type = un
;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
```

```
  class supergroup_code;
  model sqrt_eannkwh = super_mean hsize_s / solution ddfm = bw
notest;
  random intercept / subject = supergroup_code type = un ;
run;

ods rtf close;

ods rtf file="C:\Users\steve\Documents\Papers and thesis\Chapter
7 - Critical assessment\Ch7 multilevel centred group.rtf";
/* Unconditional means model */
proc mixed data = withmeans2 covtest noclprint;
   class group_code;
   model sqrt_eannkwh = / solution;
   random intercept / subject = group_code;
run;


/* Taking into account level2 means only - ecological model */
proc mixed data = withmeans2 covtest noclprint;
   class group_code;
   model sqrt_eannkwh = group_mean / solution ddfm = bw;
   random intercept / subject = group_code;
run;

/* Include effect of individual-level predictor of household size
*/
data withmeans96; set withmeans2;
hsize_s = c_rooms_hsize96 - super_mean;
hsize = c_rooms_hsize96 - group_mean;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
  class group_code;
  model sqrt_eannkwh = hsize / solution ddfm = bw notest;
  random intercept hsize / subject = group_code type = un gcorr;
run;


/* Include both effects */

proc mixed data = withmeans96 noclprint covtest noitprint;
  class group_code;
  model sqrt_eannkwh = group_mean  hsize hsize*group_mean /
solution ddfm = bw notest;
  random intercept hsize / subject = group_code type = un ;
run;

/* Test binomial predictors:
Surburban / not suburban
Urban / not urban
House / not house (flat)
Head of household over 60 / not over 60
House built before 1946 / after 1945 */


proc mixed data = withmeans96 noclprint covtest noitprint;
  class group_code;
```

```
  model sqrt_eannkwh = group_mean suburb hsize hsize*group_mean
hsize*suburb  / solution ddfm = bw notest;
  random intercept hsize / subject = group_code type = un ;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
  class group_code;
  model sqrt_eannkwh = group_mean urban hsize hsize*group_mean
hsize*urban  / solution ddfm = bw notest;
  random intercept hsize / subject = group_code type = un ;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
  class group_code;
  model sqrt_eannkwh = group_mean house hsize hsize*group_mean
hsize*house  / solution ddfm = bw notest;
  random intercept hsize / subject = group_code type = un ;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
  class group_code;
  model sqrt_eannkwh = group_mean elderly hsize hsize*group_mean
hsize*elderly  / solution ddfm = bw notest;
  random intercept hsize / subject = group_code type = un ;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
  class group_code;
  model sqrt_eannkwh = group_mean prewar hsize hsize*group_mean
hsize*prewar group_mean*prewar group_mean*hsize*prewar
/ solution ddfm = bw notest;
  random intercept hsize / subject = group_code type = un ;
run;

proc mixed data = withmeans96 noclprint covtest noitprint;
  class group_code;
  model sqrt_eannkwh = group_mean prewar hsize hsize*group_mean
hsize*prewar group_mean*prewar group_mean*hsize*prewar
/ solution ddfm = bw notest;
  random intercept  / subject = group_code ; run;


  proc mixed data = withmeans96 noclprint covtest noitprint;
  class group_code;
  model sqrt_eannkwh = group_mean house prewar hsize
hsize*group_mean hsize*prewar
/ solution ddfm = bw notest;
  random intercept  / subject = group_code ; run;

    proc mixed data = withmeans96 noclprint covtest noitprint;
  class group_code;
  model sqrt_eannkwh = group_mean prewar hsize group_mean*hsize
prewar*hsize
/ solution ddfm = bw notest;
  random intercept  / subject = group_code ; run;

ods rtf close;
```

Appendix B: Single-level modelling parameter estimates

Normality tests on the dependent variable of non-heating end-use energy (6.5)

Moments			
N	2399	**Sum Weights**	2399
Mean	3720.85403	**Sum Observations**	8926328.82
Std Deviation	2668.30621	**Variance**	7119858.02
Skewness	5.7196762	**Kurtosis**	67.292689
Uncorrected SS	5.0287E10	**Corrected SS**	1.70734E10
Coeff Variation	71.712198	**Std Error Mean**	54.4779231

Basic Statistical Measures			
Location		**Variability**	
Mean	3720.854	**Std Deviation**	2668
Median	3253.212	**Variance**	7119858
Mode	1277.500	**Range**	48864
		Interquartile Range	2242

Note: The mode displayed is the smallest of 13 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	68.30022	**Pr > |t|**	<.0001
Sign	**M**	1199.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1439400	**Pr >= |S|**	<.0001

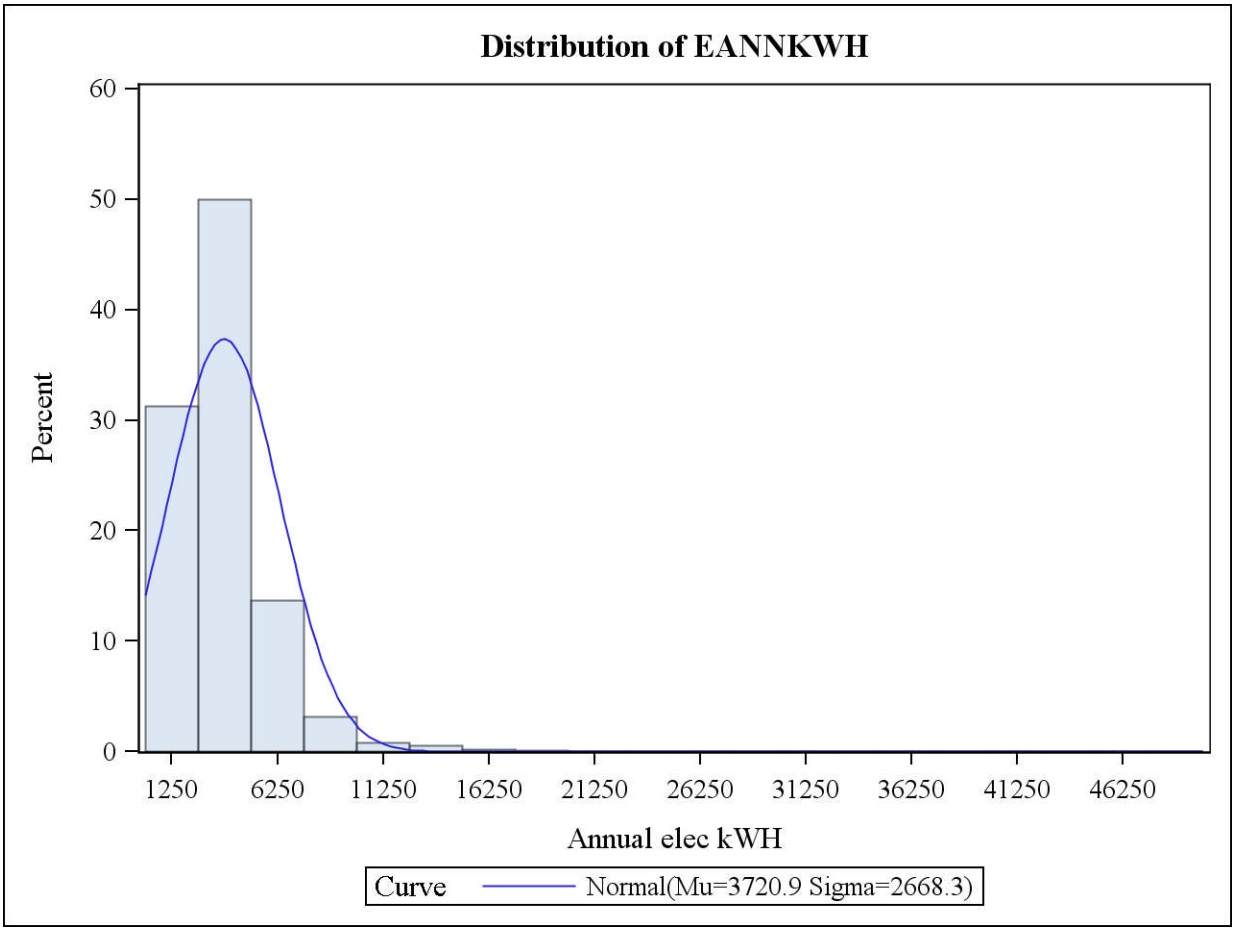
Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	48875.2109
99%	12619.2399
95%	7585.0000
90%	6140.6905
75% Q3	4519.1655
50% Median	3253.2115
25% Q1	2277.2327
10%	1526.7463
5%	1171.9487
1%	573.0701
0% Min	11.5655

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
11.5655	2011	27058.0	1064
69.9066	684	30160.5	422
94.7152	529	33405.1	807

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
141.1792	1626	41527.3	1962
192.0261	1867	48875.2	546

The UNIVARIATE Procedure

Variable: EANNKWH (Annual elec kWH)



The UNIVARIATE Procedure

Fitted Normal Distribution for EANNKWH

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	3720.854
Std Dev	Sigma	2668.306

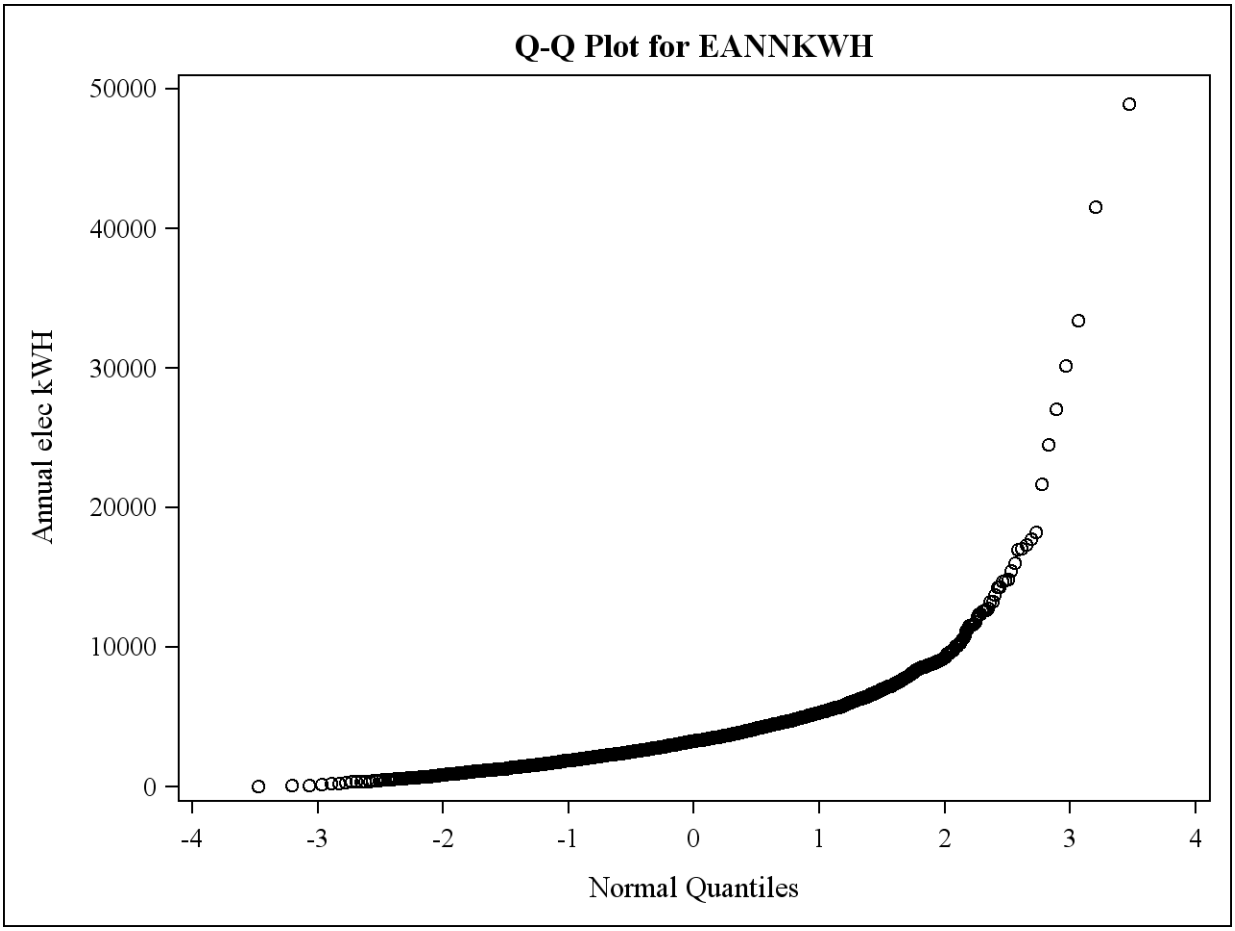
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.134410	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	19.248735	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	114.506349	**Pr > A-Sq**	<0.005

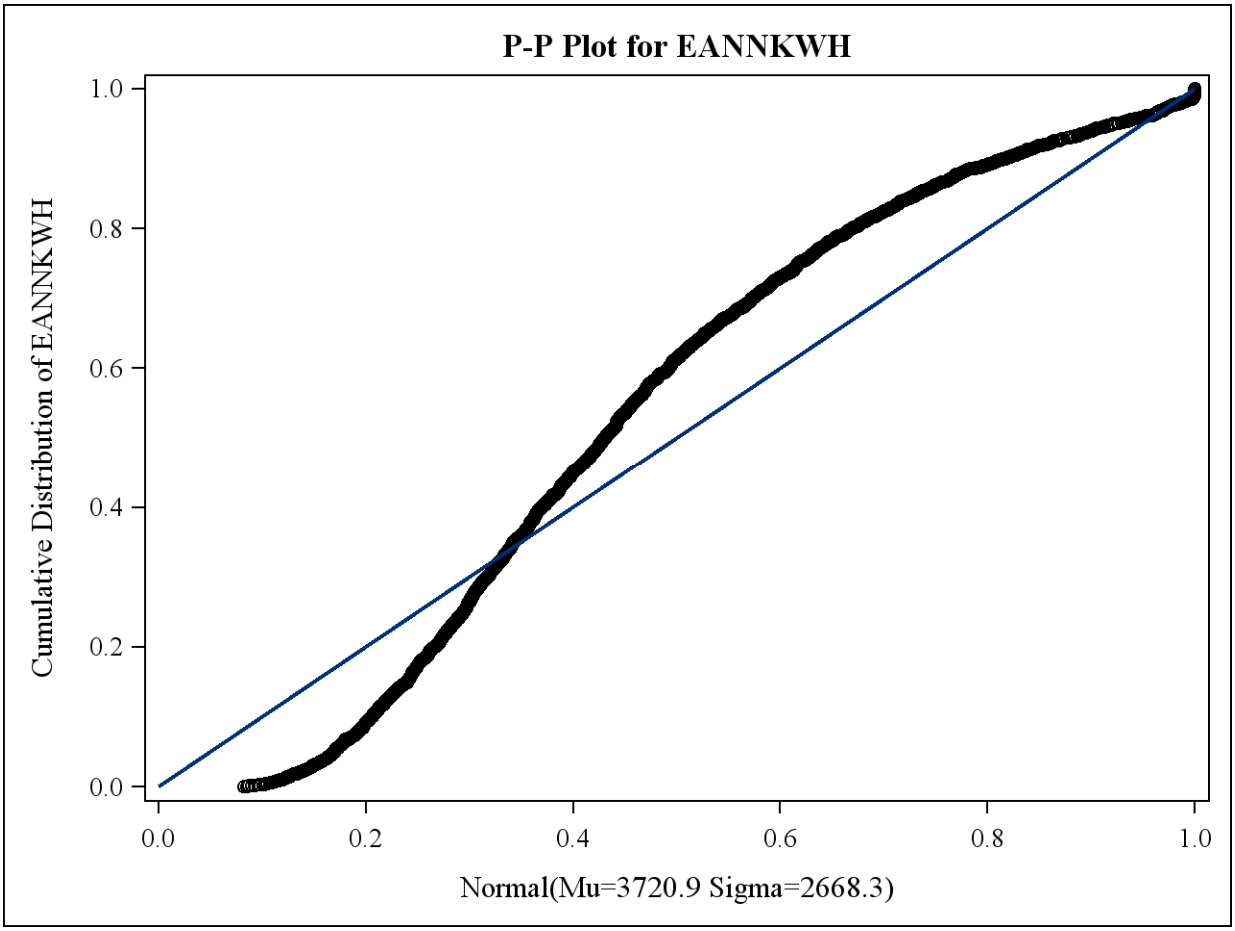
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	573.070	-2486.554
5.0	1171.949	-668.119
10.0	1526.746	301.282

The UNIVARIATE Procedure

Fitted Normal Distribution for EANNKWH

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
25.0	2277.233	1921.109
50.0	3253.212	3720.854
75.0	4519.166	5520.599
90.0	6140.690	7140.426
95.0	7585.000	8109.827
99.0	12619.240	9928.263


Q-Q Plot for EANNKWH
Annual elec kWH
50000
40000
30000
20000
10000
0
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles


P-P Plot for EANNKWH
Cumulative Distribution of EANNKWH
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Normal(Mu=3720.9 Sigma=2668.3)

The UNIVARIATE Procedure

Variable: logvar

Moments			
N	2399	**Sum Weights**	2399
Mean	8.04846095	**Sum Observations**	19308.2578
Std Deviation	0.60534044	**Variance**	0.36643705
Skewness	-0.7921115	**Kurtosis**	5.43272886
Uncorrected SS	156280.475	**Corrected SS**	878.716053
Coeff Variation	7.52119502	**Std Error Mean**	0.01235904

Basic Statistical Measures			
Location		**Variability**	
Mean	8.048461	**Std Deviation**	0.60534
Median	8.087398	**Variance**	0.36644
Mode	7.152660	**Range**	8.34900
		Interquartile Range	0.68537

The UNIVARIATE Procedure

Variable: logvar

Note: The mode displayed is the smallest of 13 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	651.2208	**Pr > |t|**	<.0001
Sign	**M**	1199.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1439400	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	10.79703
99%	9.44298
95%	8.93393
90%	8.72269
75% Q3	8.41608
50% Median	8.08740
25% Q1	7.73072

The UNIVARIATE Procedure

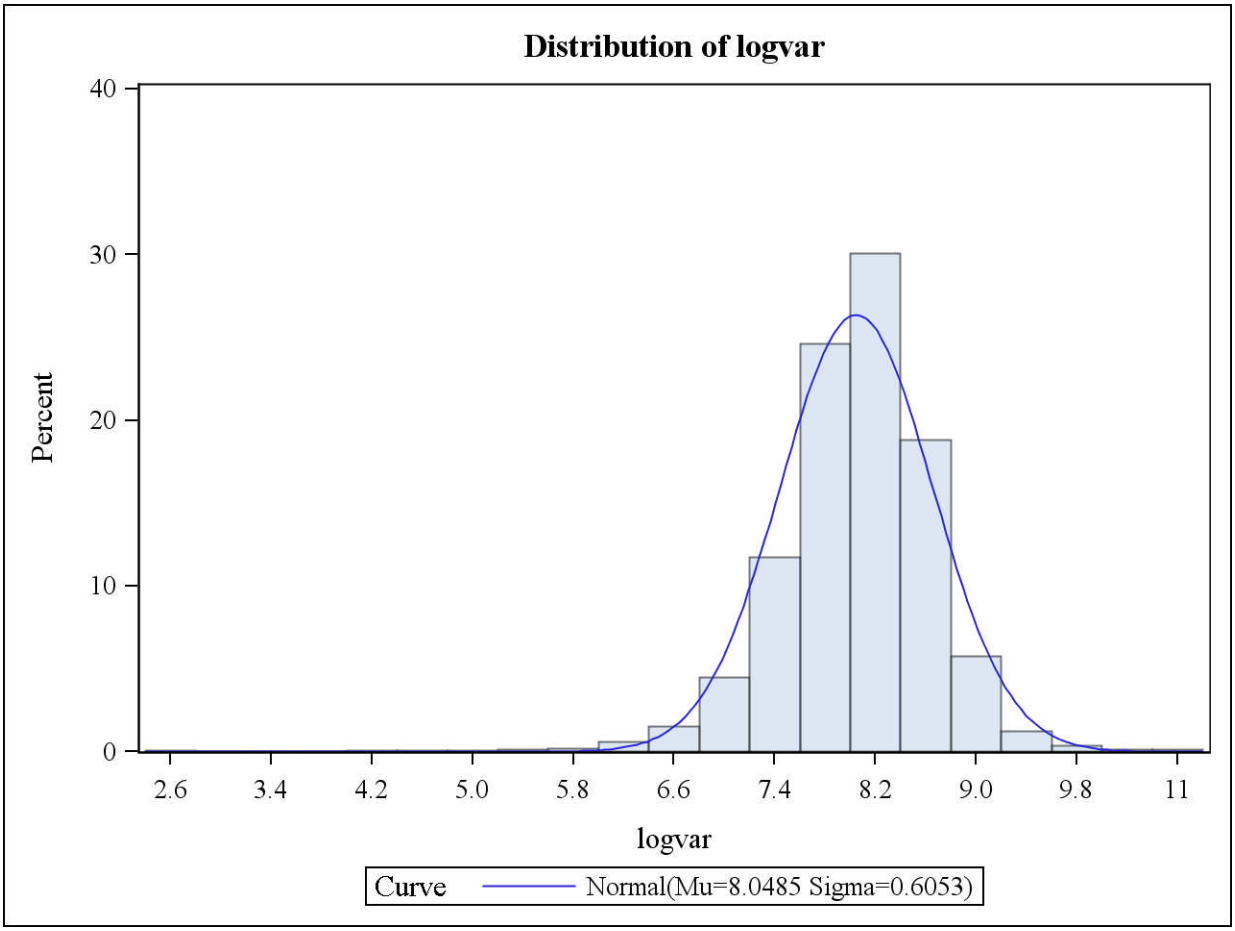
Variable: logvar

Quantiles (Definition 5)	
Quantile	**Estimate**
10%	7.33089
5%	7.06642
1%	6.35101
0% Min	2.44803

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
2.44803	2011	10.2057	1064
4.24716	684	10.3143	422
4.55087	529	10.4165	807
4.95003	1626	10.6341	1962
5.25763	1867	10.7970	546

The UNIVARIATE Procedure

Variable: logvar



The UNIVARIATE Procedure

Fitted Normal Distribution for logvar

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	8.048461
Std Dev	Sigma	0.60534

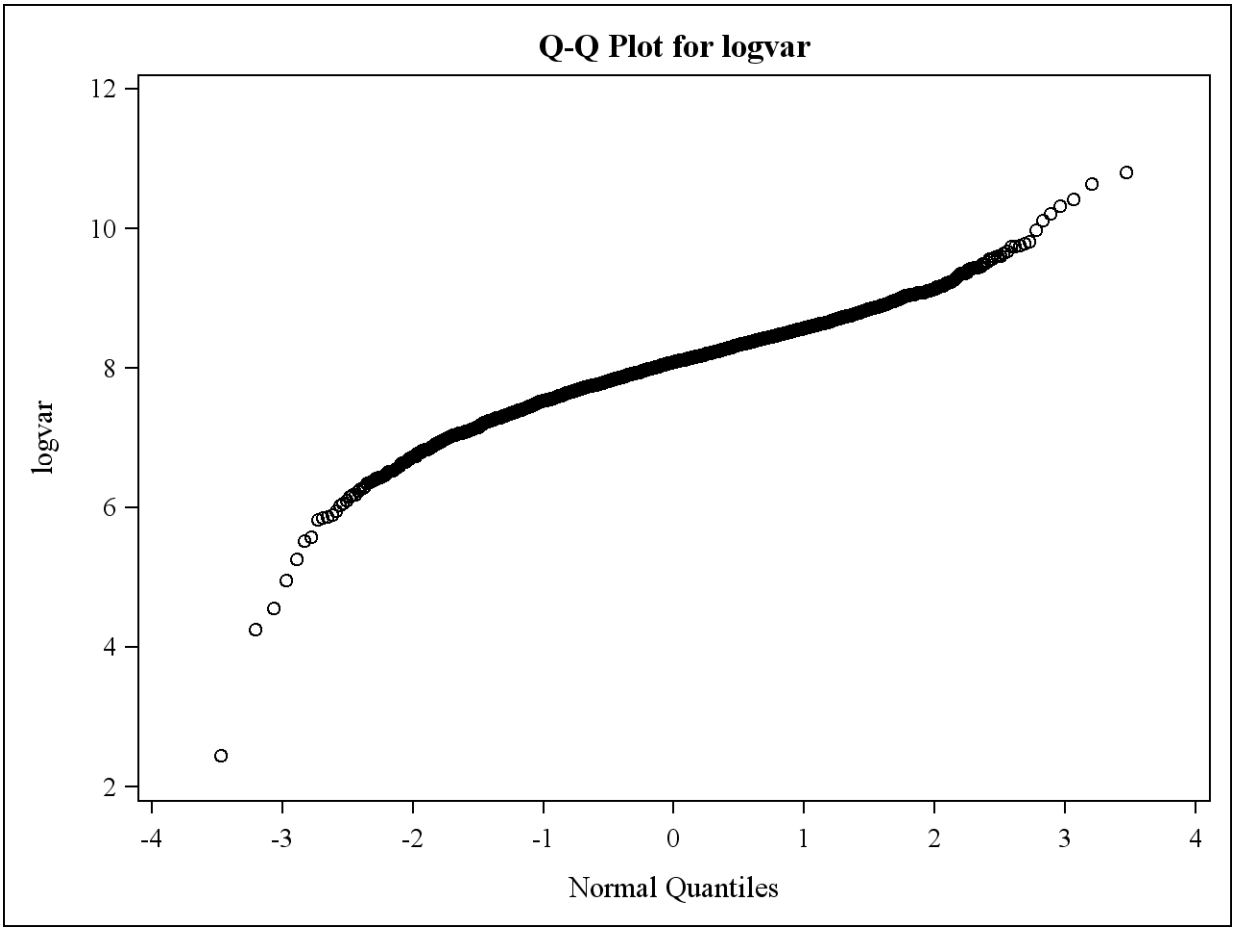
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.0507058	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	2.0580207	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	12.769756 7	**Pr > A-Sq**	<0.005

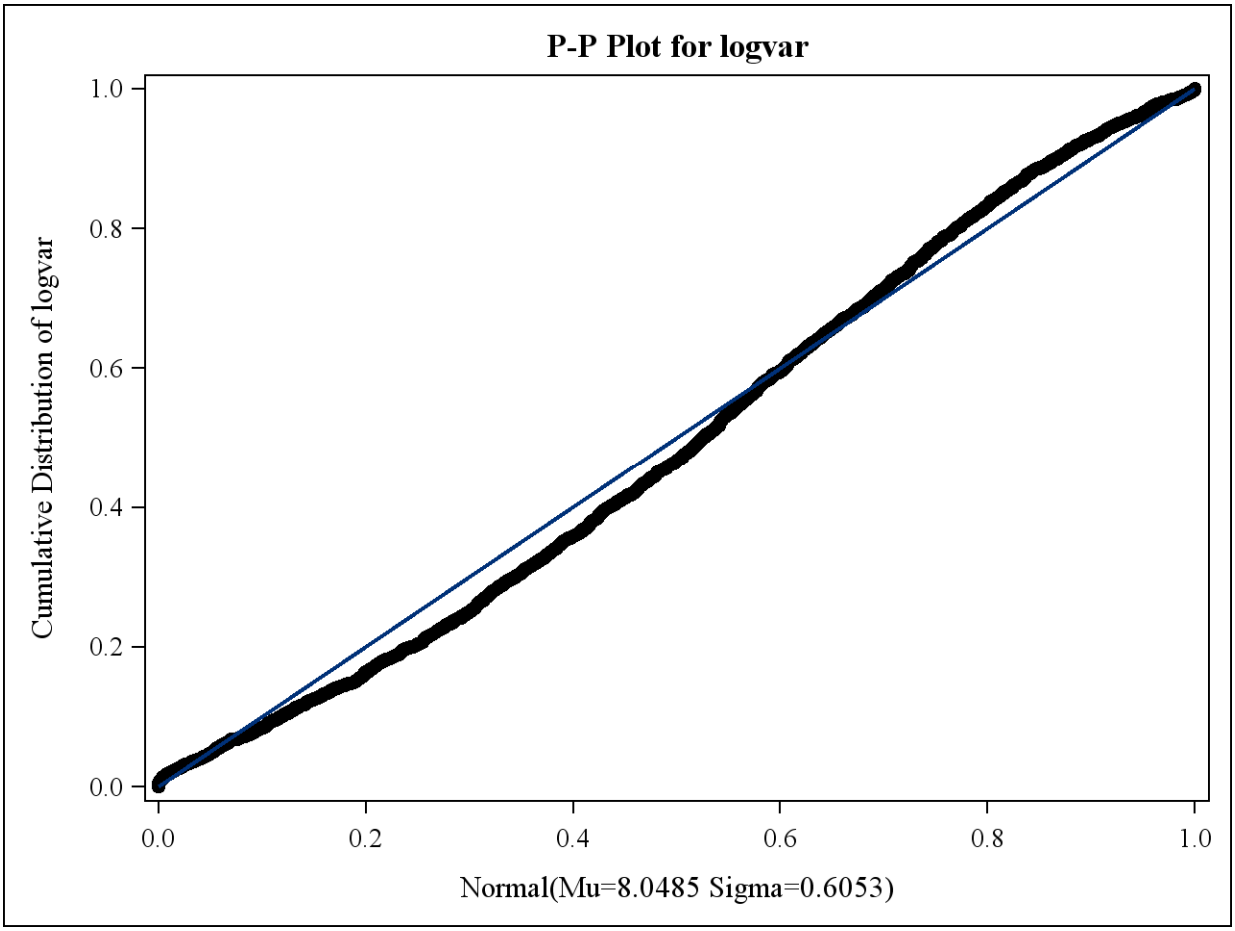
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	6.35101	6.64023
5.0	7.06642	7.05276
10.0	7.33089	7.27269

The UNIVARIATE Procedure

Fitted Normal Distribution for logvar

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
25.0	7.73072	7.64017
50.0	8.08740	8.04846
75.0	8.41608	8.45676
90.0	8.72269	8.82424
95.0	8.93393	9.04416
99.0	9.44298	9.45669


Q-Q Plot for logvar
logvar
12
10
8
6
4
2
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles


P-P Plot for logvar
Cumulative Distribution of logvar
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Normal(Mu=8.0485 Sigma=0.6053)

The UNIVARIATE Procedure

Variable: sqrtvar

Moments			
N	2399	**Sum Weights**	2399
Mean	58.4281164	**Sum Observations**	140169.051
Std Deviation	17.5253322	**Variance**	307.137268
Skewness	1.53810007	**Kurtosis**	8.54552522
Uncorrected SS	8926328.82	**Corrected SS**	736515.169
Coeff Variation	29.9946896	**Std Error Mean**	0.3578089

Basic Statistical Measures			
Location		**Variability**	
Mean	58.42812	**Std Deviation**	17.52533
Median	57.03693	**Variance**	307.13727
Mode	35.74213	**Range**	217.67657
		Interquartile Range	19.50438

The UNIVARIATE Procedure

Variable: sqrtvar

Note: The mode displayed is the smallest of 13 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	163.2942	**Pr > |t|**	<.0001
Sign	**M**	1199.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1439400	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	221.07739
99%	112.33539
95%	87.09191
90%	78.36256
75% Q3	67.22474
50% Median	57.03693
25% Q1	47.72036
10%	39.07360

The UNIVARIATE Procedure

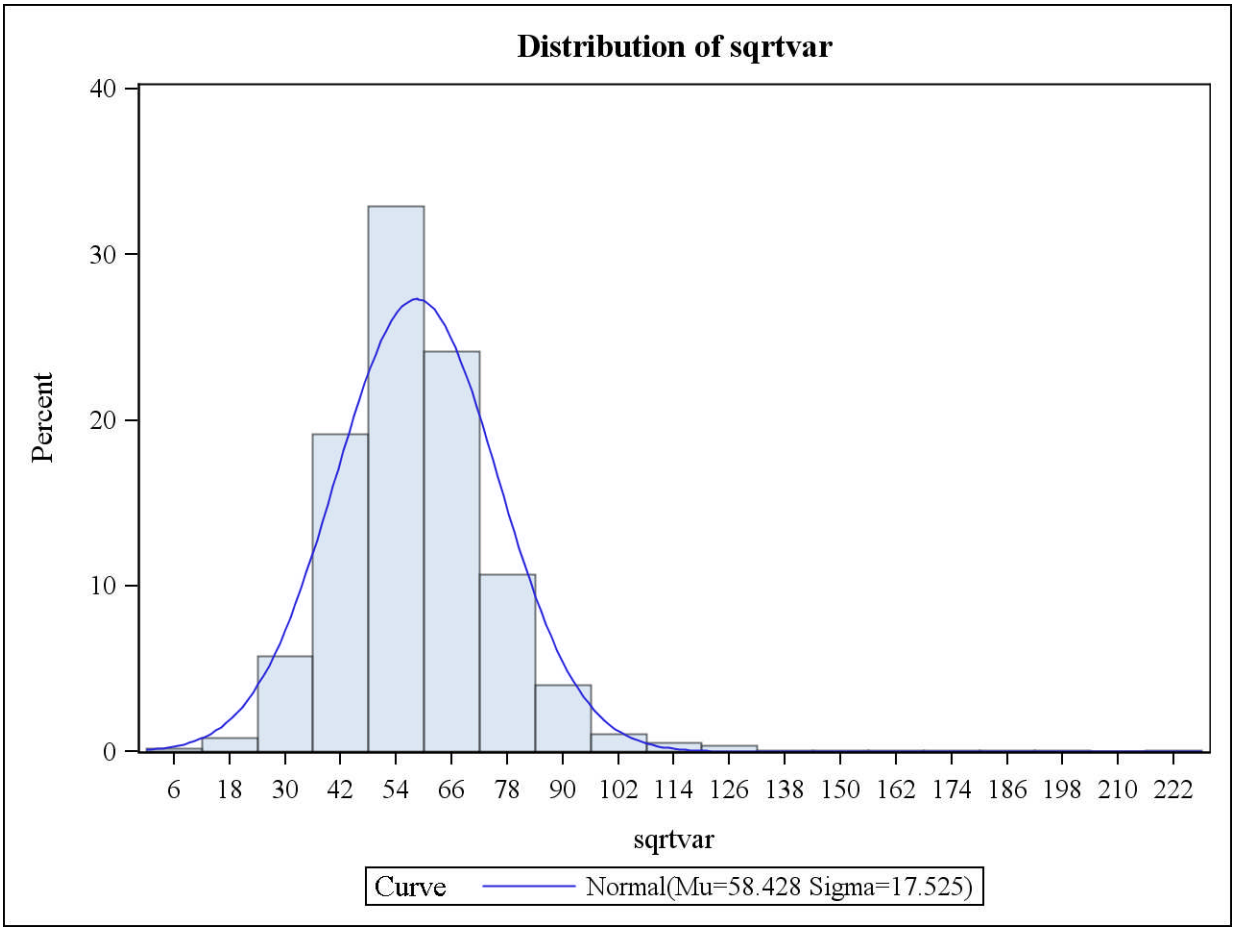
Variable: sqrtvar

Quantiles (Definition 5)	
Quantile	**Estimate**
5%	34.23374
1%	23.93888
0% Min	3.40081

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
3.40081	2011	164.493	1064
8.36101	684	173.668	422
9.73217	529	182.771	807
11.88189	1626	203.782	1962
13.85735	1867	221.077	546

The UNIVARIATE Procedure

Variable: sqrtvar



The UNIVARIATE Procedure

Fitted Normal Distribution for sqrtvar

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	58.42812
Std Dev	Sigma	17.52533

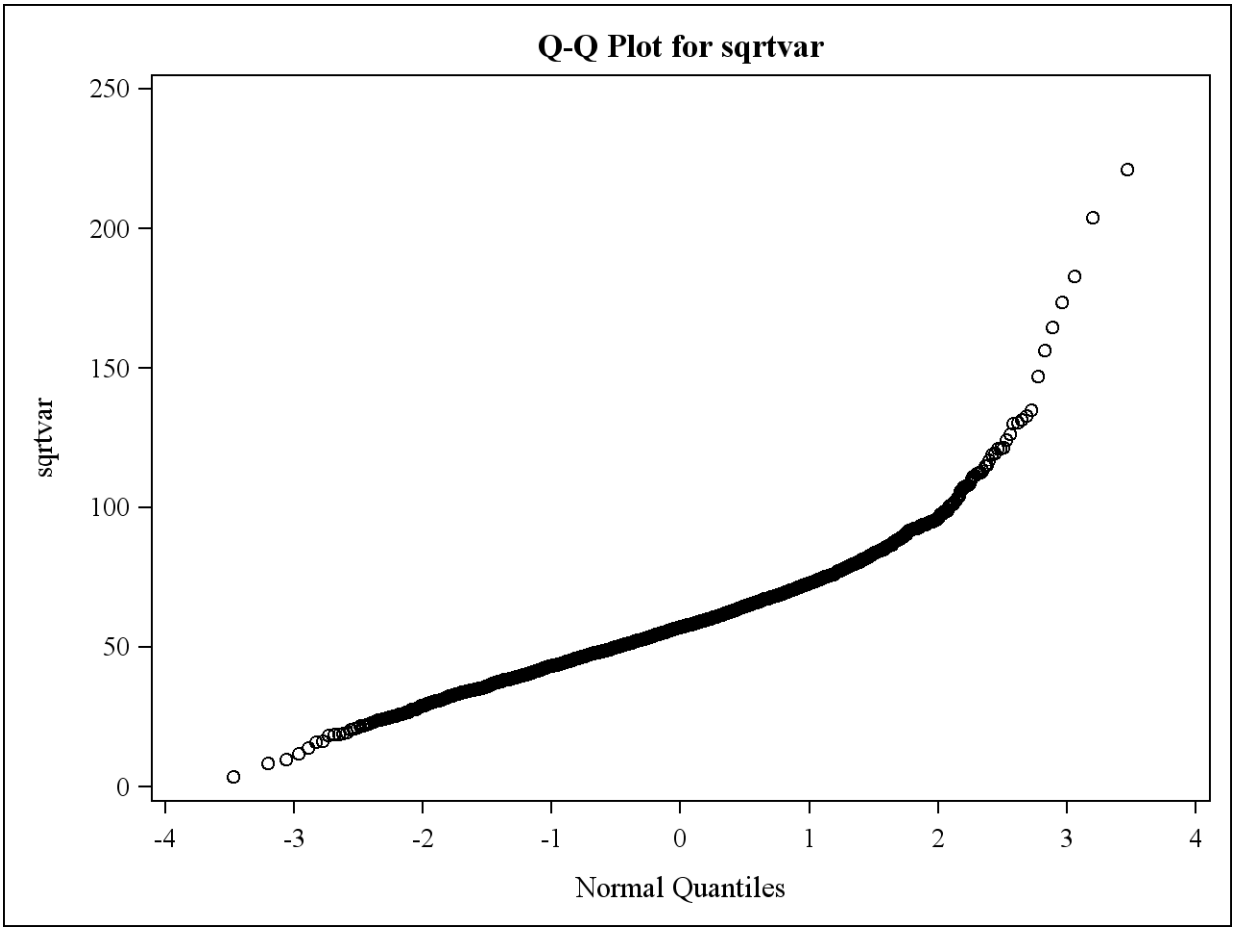
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.0592638	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	3.2466195	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	20.3956338	**Pr > A-Sq**	<0.005

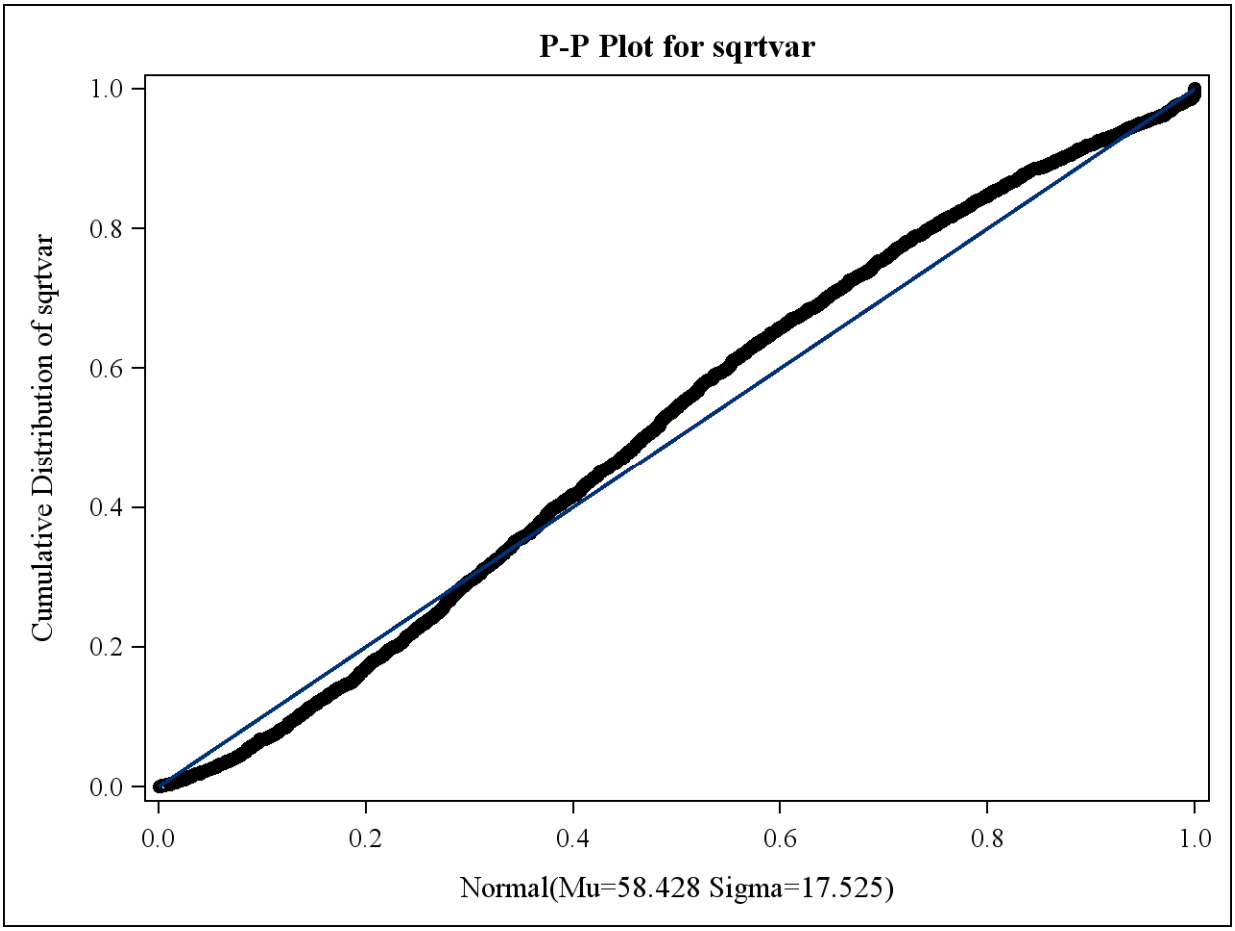
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	23.9389	17.6581
5.0	34.2337	29.6015
10.0	39.0736	35.9685
25.0	47.7204	46.6075

The UNIVARIATE Procedure

Fitted Normal Distribution for sqrtvar

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
50.0	57.0369	58.4281
75.0	67.2247	70.2488
90.0	78.3626	80.8877
95.0	87.0919	87.2547
99.0	112.3354	99.1981


Q-Q Plot for sqrtvar
sqrtvar
250
200
150
100
50
0
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles


P-P Plot for sqrtvar
Cumulative Distribution of sqrtvar
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Normal(Mu=58.428 Sigma=17.525)

The UNIVARIATE Procedure

Variable: fthrtvar

Moments			
N	2399	**Sum Weights**	2399
Mean	7.56253238	**Sum Observations**	18142.5152
Std Deviation	1.11208628	**Variance**	1.2367359
Skewness	0.39647743	**Kurtosis**	3.065272
Uncorrected SS	140169.051	**Corrected SS**	2965.69269
Coeff Variation	14.7052102	**Std Error Mean**	0.0227051

Basic Statistical Measures			
Location		**Variability**	
Mean	7.562532	**Std Deviation**	1.11209
Median	7.552280	**Variance**	1.23674
Mode	5.978472	**Range**	13.02454
		Interquartile Range	1.29108

Note: The mode displayed is the smallest of 13 modes with a count of 2.

The UNIVARIATE Procedure

Variable: fthrtvar

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	333.0764	**Pr > |t|**	<.0001
Sign	**M**	1199.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1439400	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	14.86867
99%	10.59884
95%	9.33230
90%	8.85226
75% Q3	8.19907
50% Median	7.55228
25% Q1	6.90799
10%	6.25089
5%	5.85096
1%	4.89274
0% Min	1.84413

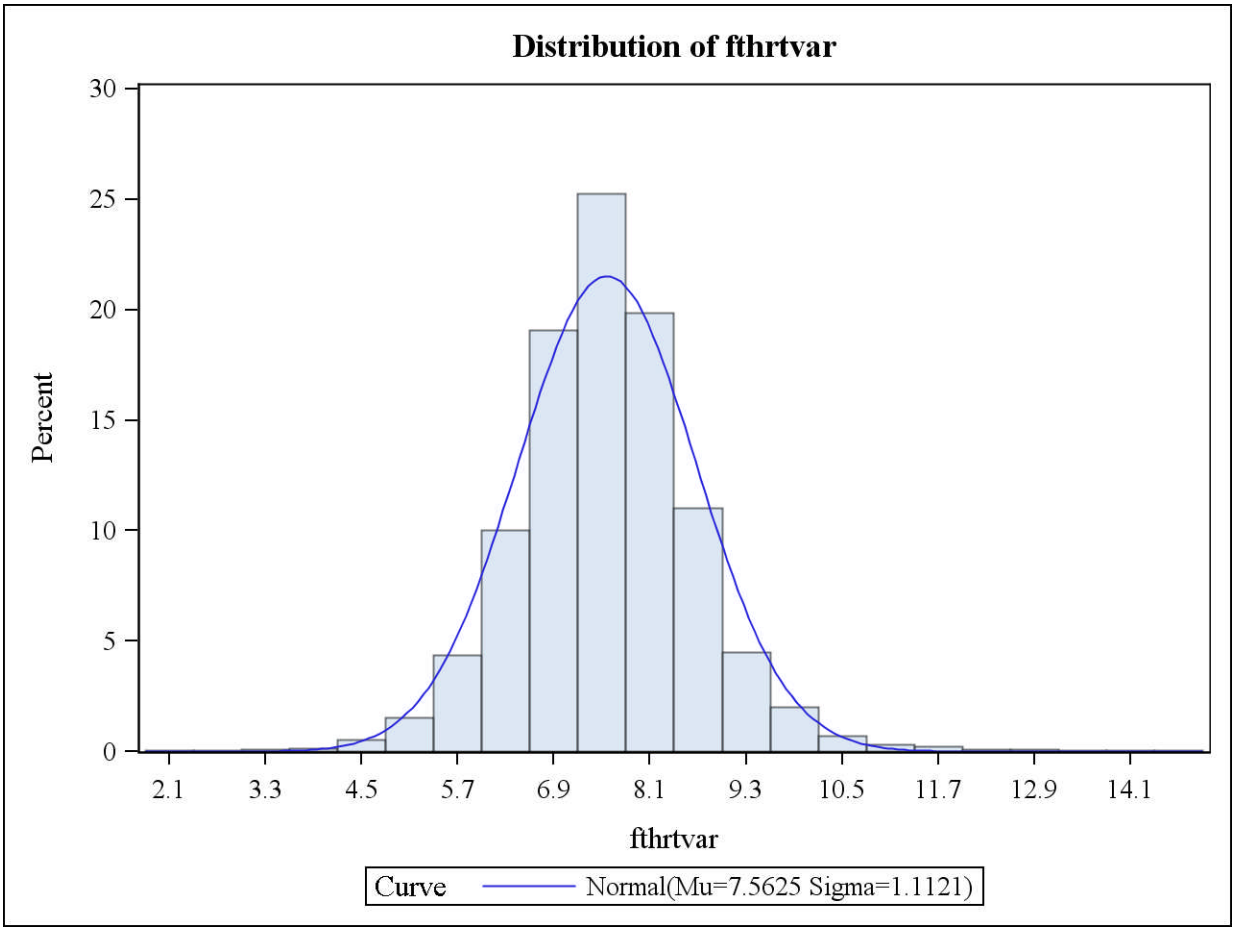
The UNIVARIATE Procedure

Variable: fthrtvar

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.84413	2011	12.8255	1064
2.89154	684	13.1783	422
3.11964	529	13.5193	807
3.44701	1626	14.2752	1962
3.72255	1867	14.8687	546

The UNIVARIATE Procedure

Variable: fthrtvar



The UNIVARIATE Procedure

Fitted Normal Distribution for fthrtvar

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	7.562532
Std Dev	Sigma	1.112086

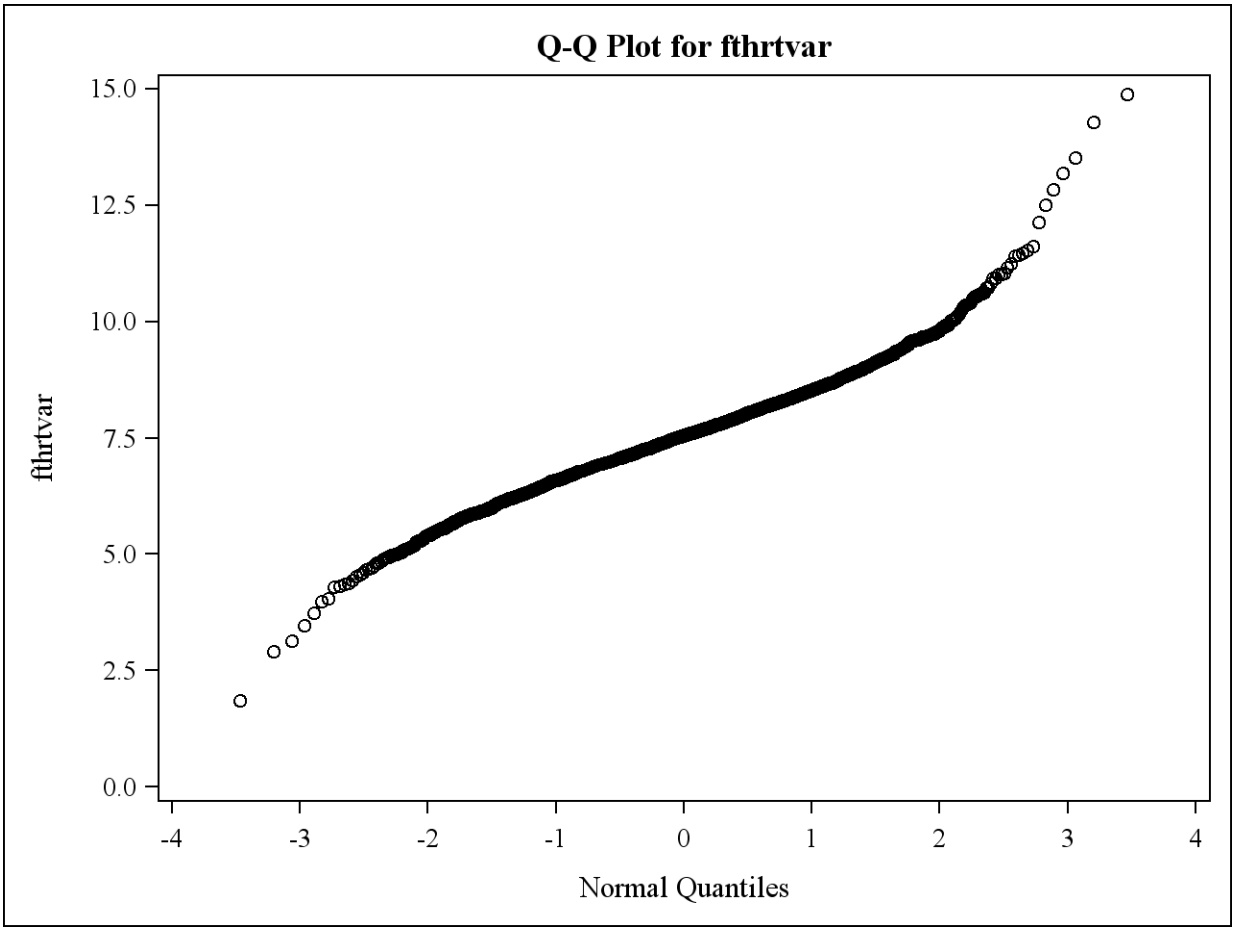
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.03735320	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	1.17296184	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	7.67559742	**Pr > A-Sq**	<0.005

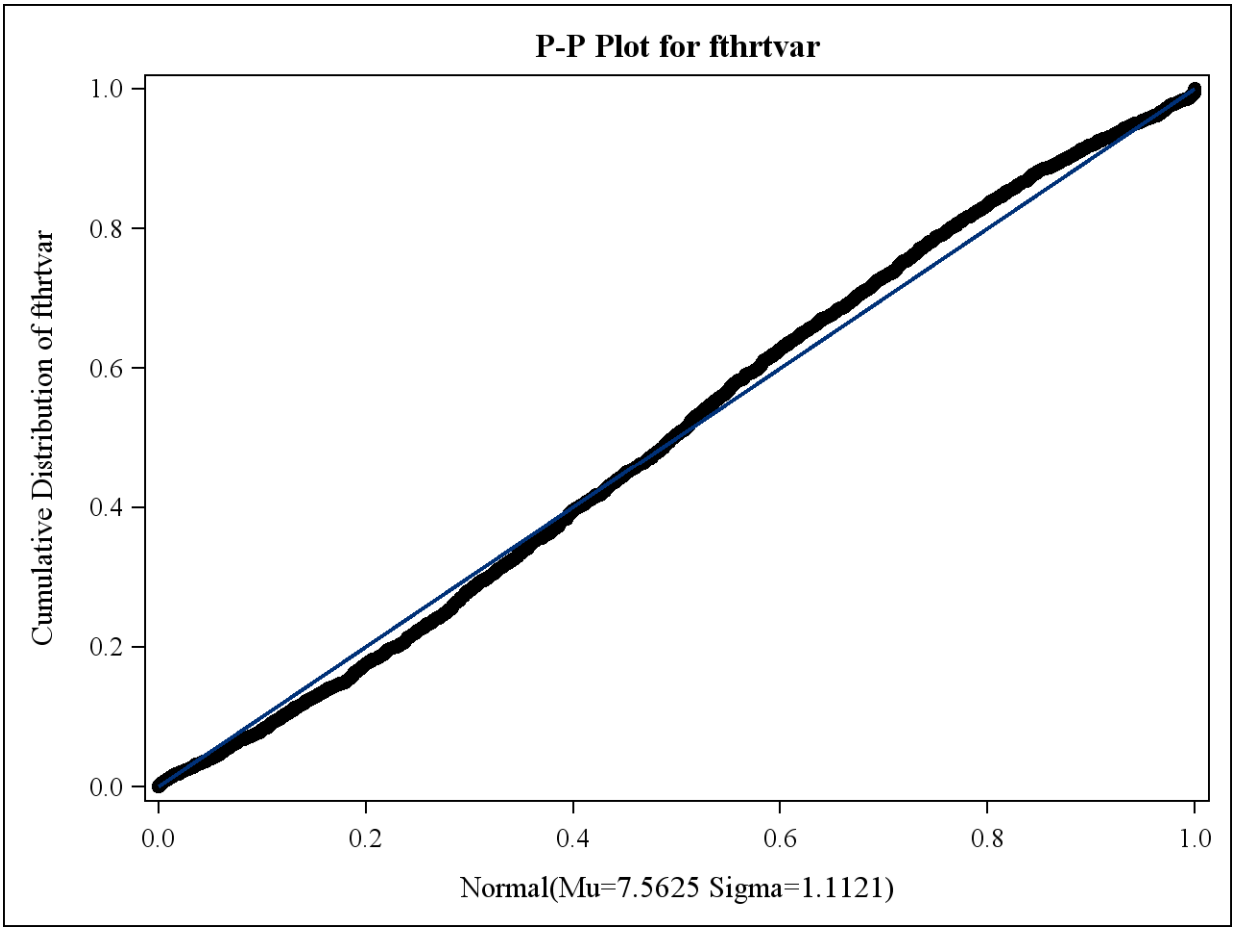
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	4.89274	4.97543
5.0	5.85096	5.73331
10.0	6.25089	6.13734
25.0	6.90799	6.81244
50.0	7.55228	7.56253

The UNIVARIATE Procedure

Fitted Normal Distribution for fthrtvar

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
75.0	8.19907	8.31262
90.0	8.85226	8.98773
95.0	9.33230	9.39175
99.0	10.59884	10.14963


Q-Q Plot for fthrtvar
fthrtvar
15.0
12.5
10.0
7.5
5.0
2.5
0.0
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles


P-P Plot for fthrtvar
Cumulative Distribution of fthrtvar
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Normal(Mu=7.5625 Sigma=1.1121)

The UNIVARIATE Procedure

Variable: sqvar

Moments			
N	2399	**Sum Weights**	2399
Mean	20961644.9	**Sum Observations**	5.0287E10
Std Deviation	74257153	**Variance**	5.51412E15
Skewness	21.5087699	**Kurtosis**	579.639972
Uncorrected SS	1.4277E19	**Corrected SS**	1.32229E19
Coeff Variation	354.252509	**Std Error Mean**	1516083.67

Basic Statistical Measures			
Location		**Variability**	
Mean	20961645	**Std Deviation**	74257153
Median	10583385	**Variance**	5.51412E15
Mode	1632006	**Range**	2388786110
		Interquartile Range	15237068

Note: The mode displayed is the smallest of 13 modes with a count of 2.

The UNIVARIATE Procedure

Variable: sqvar

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	13.82618	**Pr > |t|**	<.0001
Sign	**M**	1199.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1439400	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	2.38879E+09
99%	1.59245E+08
95%	5.75322E+07
90%	3.77081E+07
75% Q3	2.04229E+07
50% Median	1.05834E+07
25% Q1	5.18579E+06
10%	2.33095E+06

The UNIVARIATE Procedure

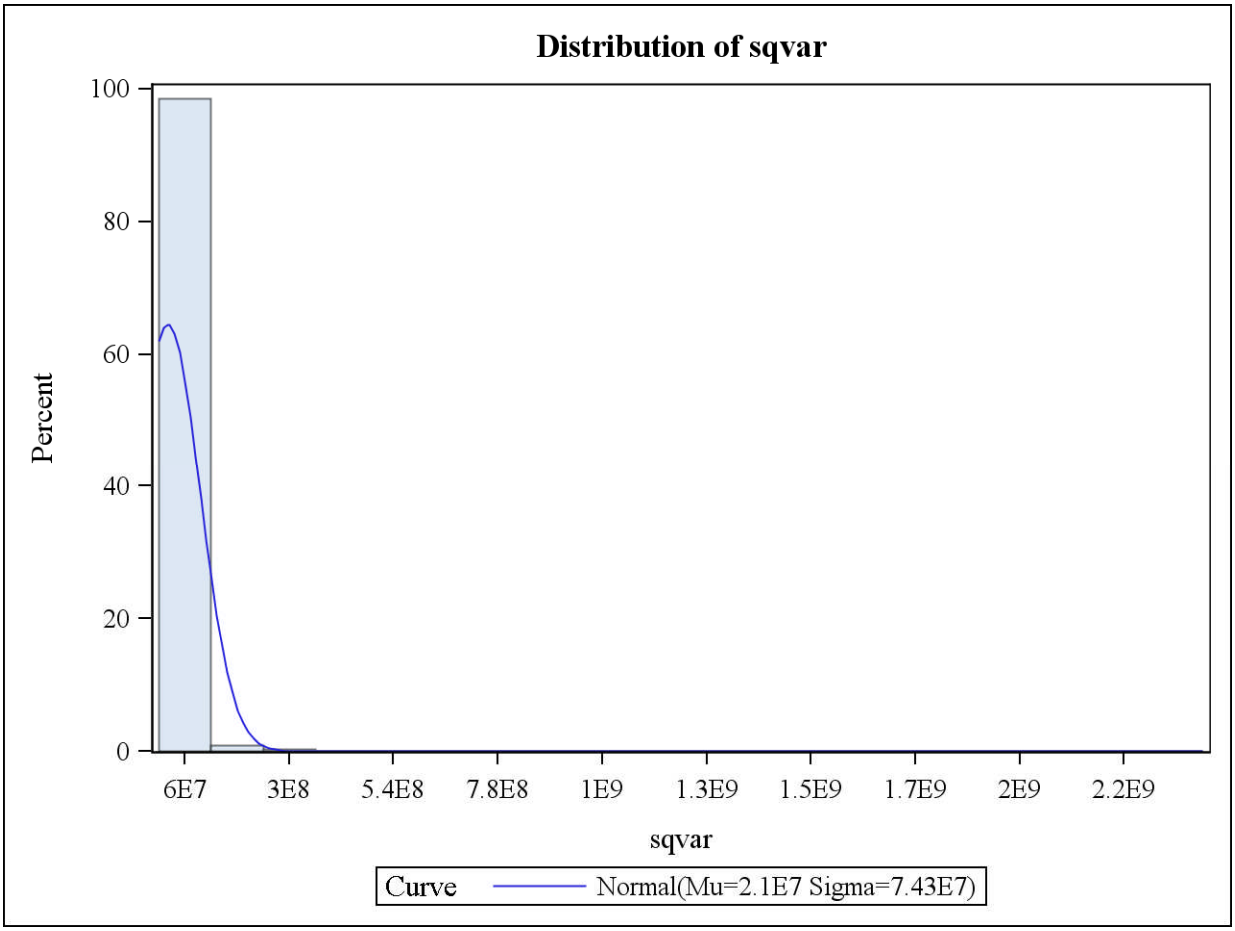
Variable: sqvar

Quantiles (Definition 5)	
Quantile	**Estimate**
5%	1.37346E+06
1%	3.28409E+05
0% Min	1.33761E+02

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
133.761	2011	732134348	1064
4886.926	684	909654037	422
8970.967	529	1115904027	807
19931.579	1626	1724514432	1962
36874.023	1867	2388786244	546

The UNIVARIATE Procedure

Variable: sqvar



The UNIVARIATE Procedure

Fitted Normal Distribution for sqvar

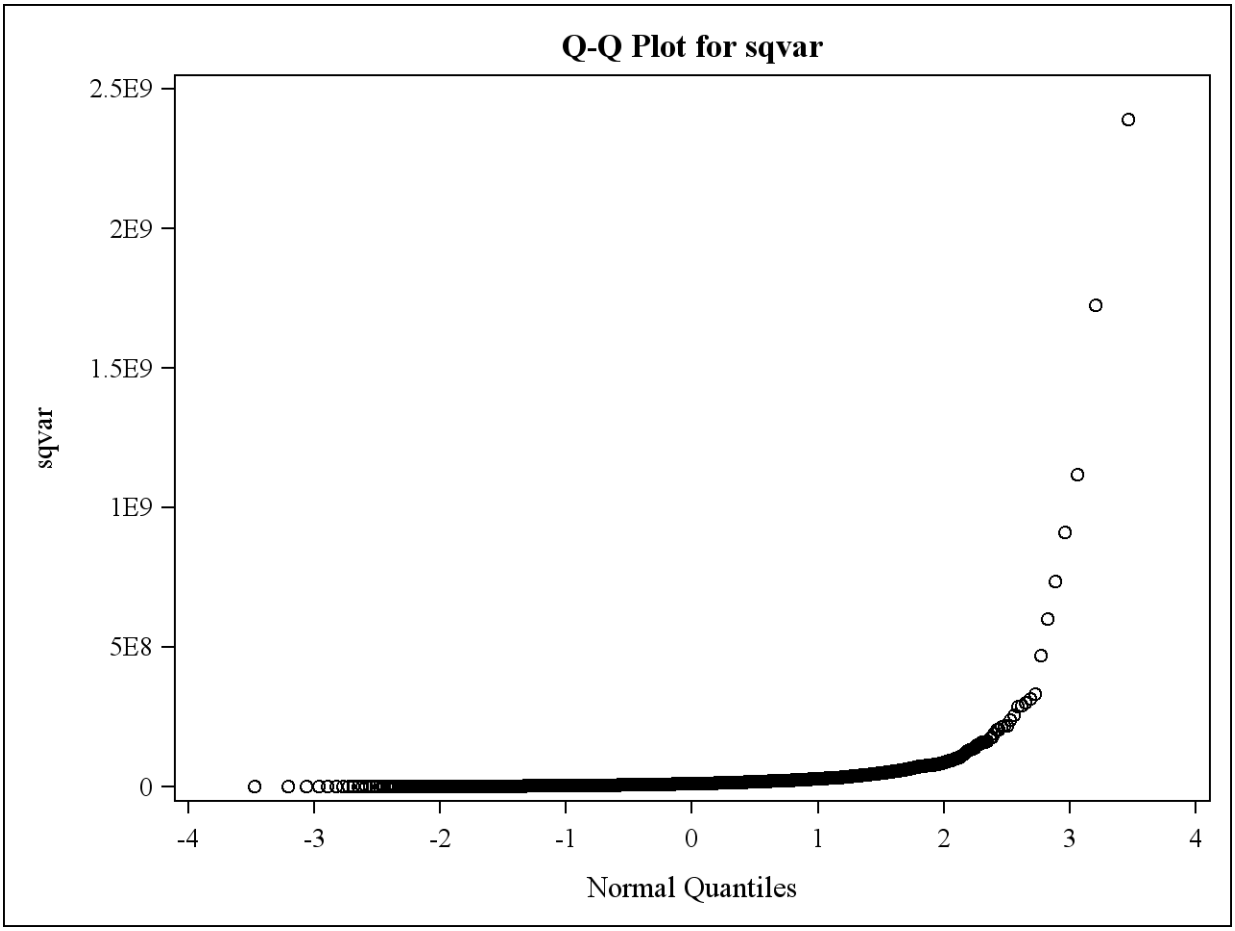
Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	20961645
Std Dev	Sigma	74257153

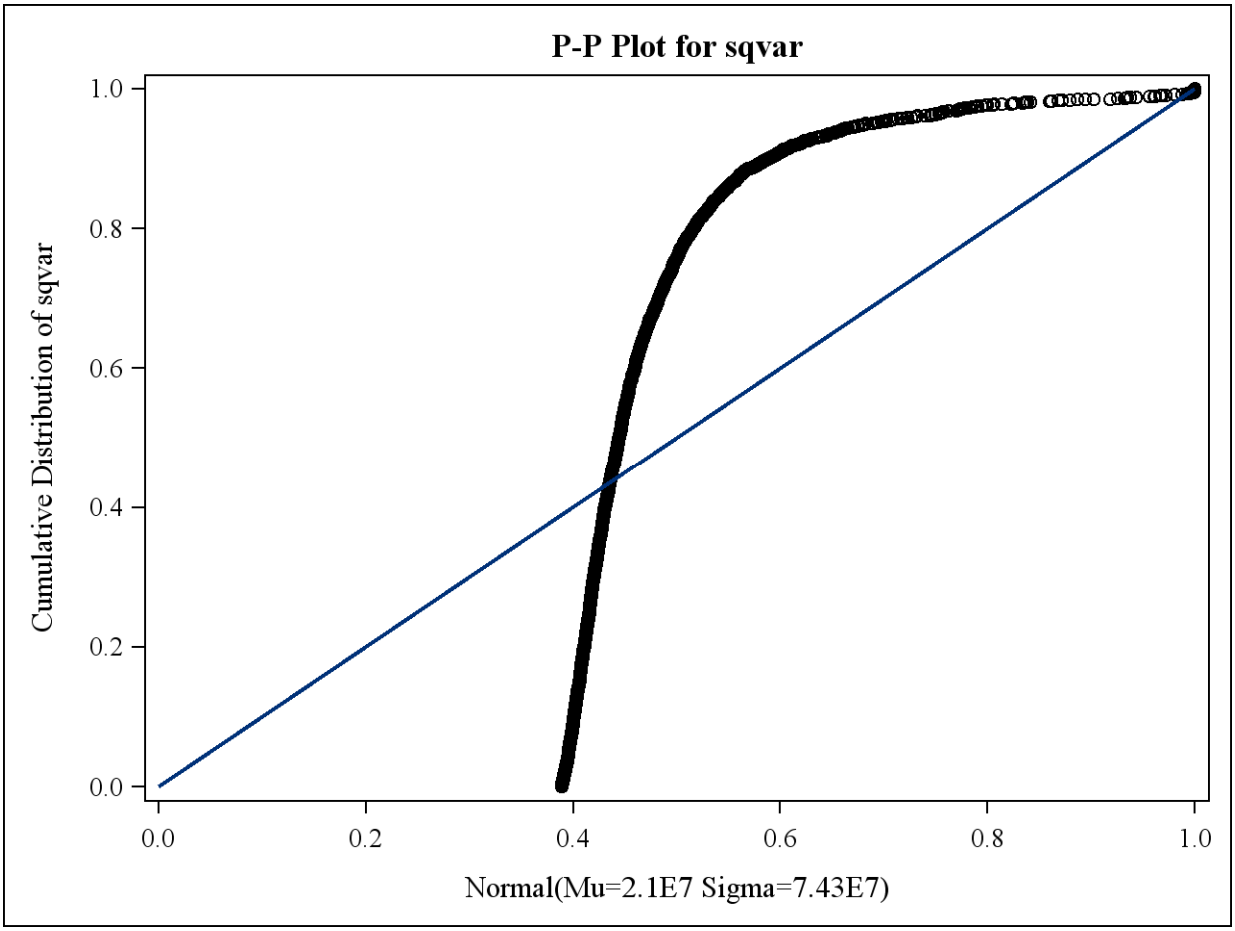
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.388863	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	115.038402	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	567.799570	**Pr > A-Sq**	<0.005

The UNIVARIATE Procedure

Fitted Normal Distribution for sqvar

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	328409	-151786325
5.0	1373464	-101180503
10.0	2330954	-74202726
25.0	5185789	-29124044
50.0	10583385	20961645
75.0	20422857	71047333
90.0	37708080	116126016
95.0	57532225	143103792
99.0	159245216	193709615


Q-Q Plot for sqvar
2.5E9
2E9
1.5E9
1E9
5E8
0
sqvar
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles


P-P Plot for sqvar
Cumulative Distribution of sqvar
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Normal(Mu=2.1E7 Sigma=7.43E7)

The UNIVARIATE Procedure

Variable: EANNKWH (Annual elec kWH)

Moments			
N	2293	**Sum Weights**	2293
Mean	3466.99572	**Sum Observations**	7949821.19
Std Deviation	1752.12936	**Variance**	3069957.29
Skewness	0.98857714	**Kurtosis**	1.20869654
Uncorrected SS	3.45983E10	**Corrected SS**	7036342109
Coeff Variation	50.537396 1	**Std Error Mean**	36.590145 7

Basic Statistical Measures			
Location		**Variability**	
Mean	3466.996	**Std Deviation**	1752
Median	3199.623	**Variance**	3069957
Mode	2021.538	**Range**	10509
		Interquartile Range	2143

The UNIVARIATE Procedure

Variable: EANNKWH (Annual elec kWH)

Note: The mode displayed is the smallest of 12 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	94.75217	**Pr > |t|**	<.0001
Sign	**M**	1146.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1315036	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	10520.5882
99%	8922.4451
95%	6912.5473
90%	5738.2212
75% Q3	4389.6319
50% Median	3199.6232
25% Q1	2246.4988
10%	1512.1429

The UNIVARIATE Procedure

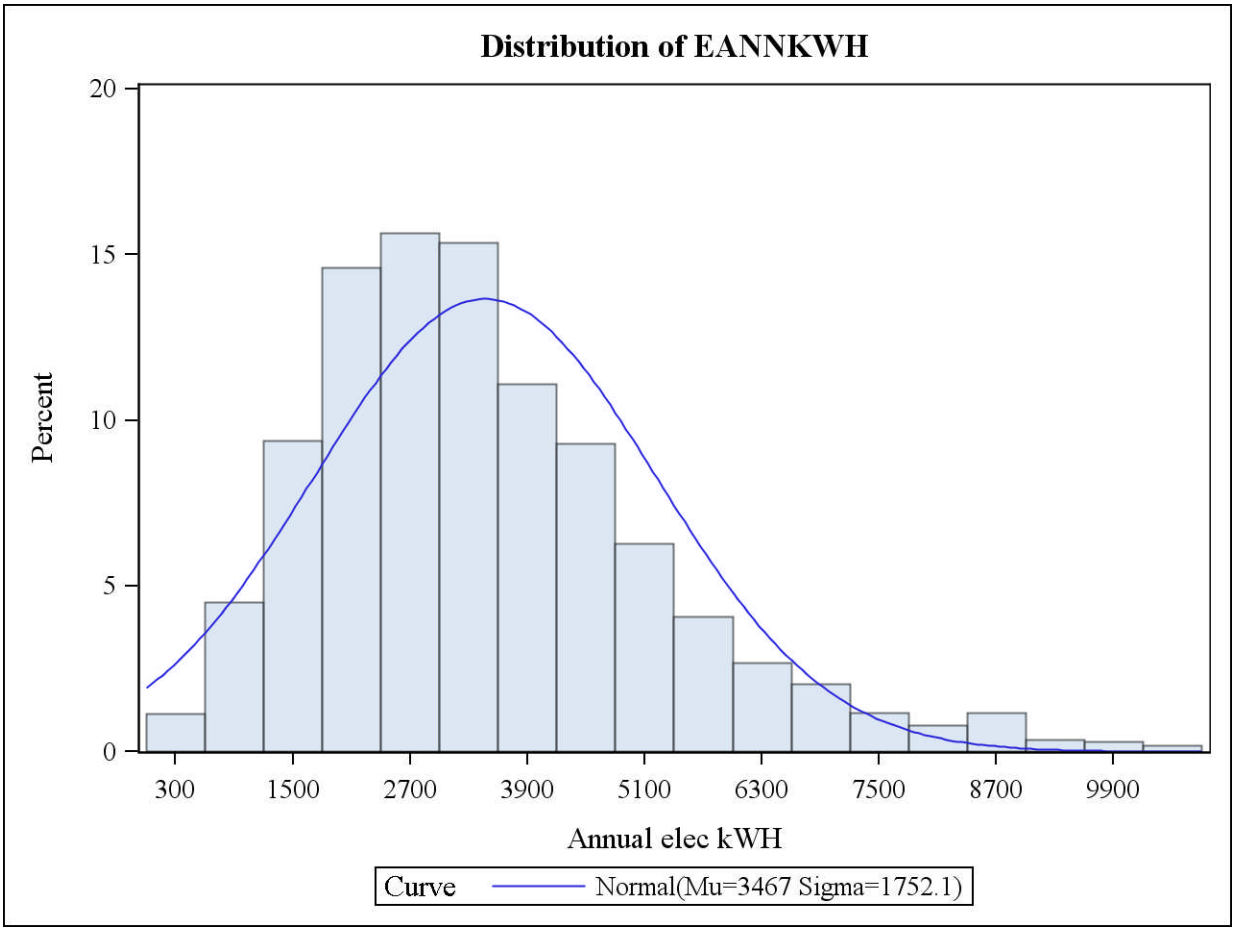
Variable: EANNKWH (Annual elec kWH)

Quantiles (Definition 5)	
Quantile	**Estimate**
5%	1160.8948
1%	567.7778
0% Min	11.5655

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
11.5655	1921	10093.2	1681
69.9066	647	10201.7	1456
94.7152	508	10268.2	256
141.1792	1547	10393.0	1834
192.0261	1781	10520.6	1860

The UNIVARIATE Procedure

Variable: EANNKWH (Annual elec kWH)



The UNIVARIATE Procedure

Fitted Normal Distribution for EANNKWH

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	3466.996
Std Dev	Sigma	1752.129

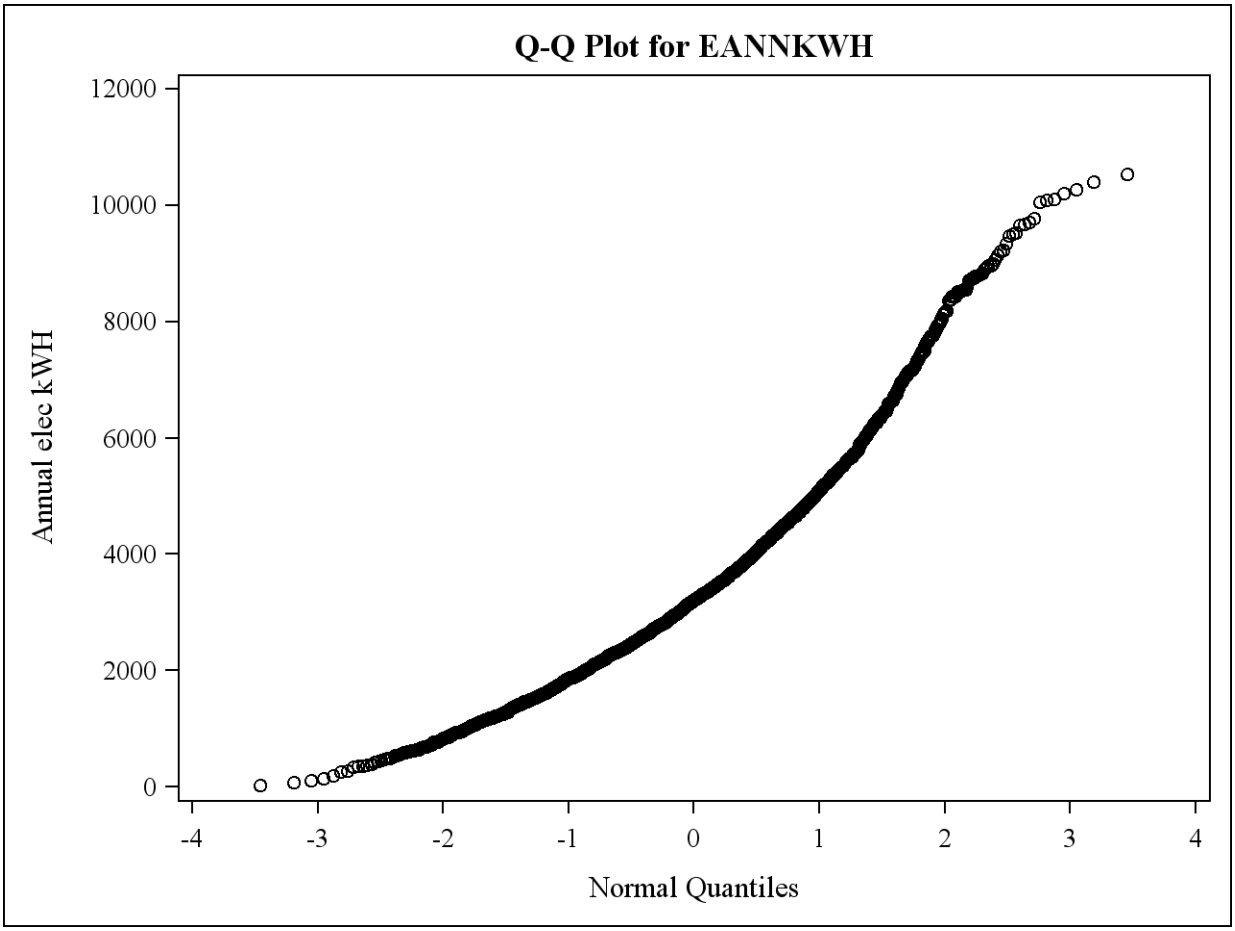
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.0782594	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	4.5732336	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	28.8050304	**Pr > A-Sq**	<0.005

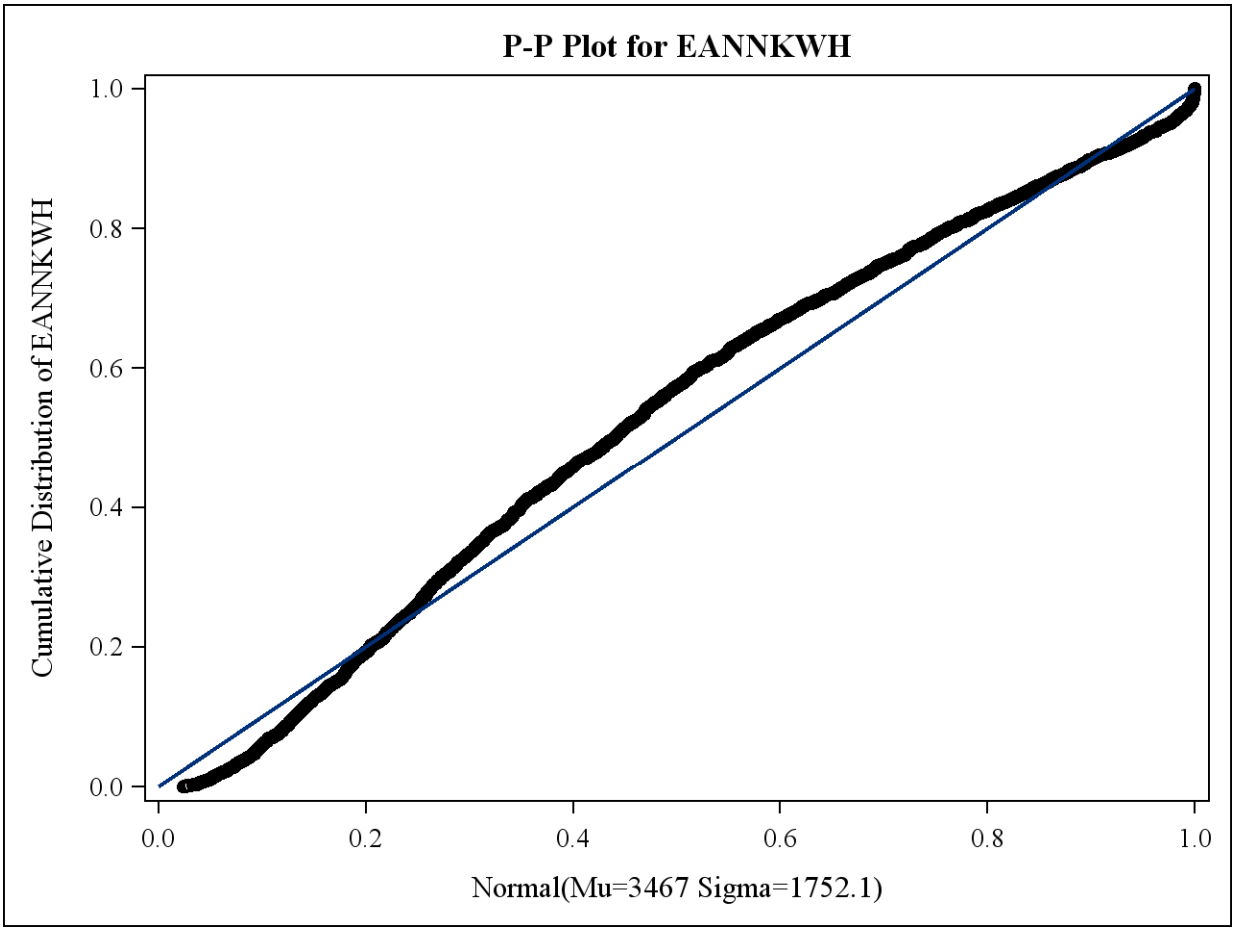
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	567.778	-609.067
5.0	1160.895	584.999
10.0	1512.143	1221.552

The UNIVARIATE Procedure

Fitted Normal Distribution for EANNKWH

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
25.0	2246.499	2285.202
50.0	3199.623	3466.996
75.0	4389.632	4648.789
90.0	5738.221	5712.440
95.0	6912.547	6348.992
99.0	8922.445	7543.058


Q-Q Plot for EANNKWH
Annual elec kWH
12000
10000
8000
6000
4000
2000
0
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles


P-P Plot for EANNKWH
Cumulative Distribution of EANNKWH
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Normal(Mu=3467 Sigma=1752.1)

The UNIVARIATE Procedure

Variable: logvar

Moments			
N	2293	**Sum Weights**	2293
Mean	8.01227882	**Sum Observations**	18372.1553
Std Deviation	0.57088997	**Variance**	0.32591536
Skewness	-1.2676943	**Kurtosis**	6.24865591
Uncorrected SS	147949.829	**Corrected SS**	746.997995
Coeff Variation	7.12518848	**Std Error Mean**	0.01192203

Basic Statistical Measures			
Location		**Variability**	
Mean	8.012279	**Std Deviation**	0.57089
Median	8.070788	**Variance**	0.32592
Mode	7.611614	**Range**	6.81306
		Interquartile Range	0.66987

The UNIVARIATE Procedure

Variable: logvar

Note: The mode displayed is the smallest of 12 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	672.0563	**Pr > |t|**	<.0001
Sign	**M**	1146.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1315036	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	9.26109
99%	9.09633
95%	8.84109
90%	8.65490
75% Q3	8.38700
50% Median	8.07079
25% Q1	7.71713

The UNIVARIATE Procedure

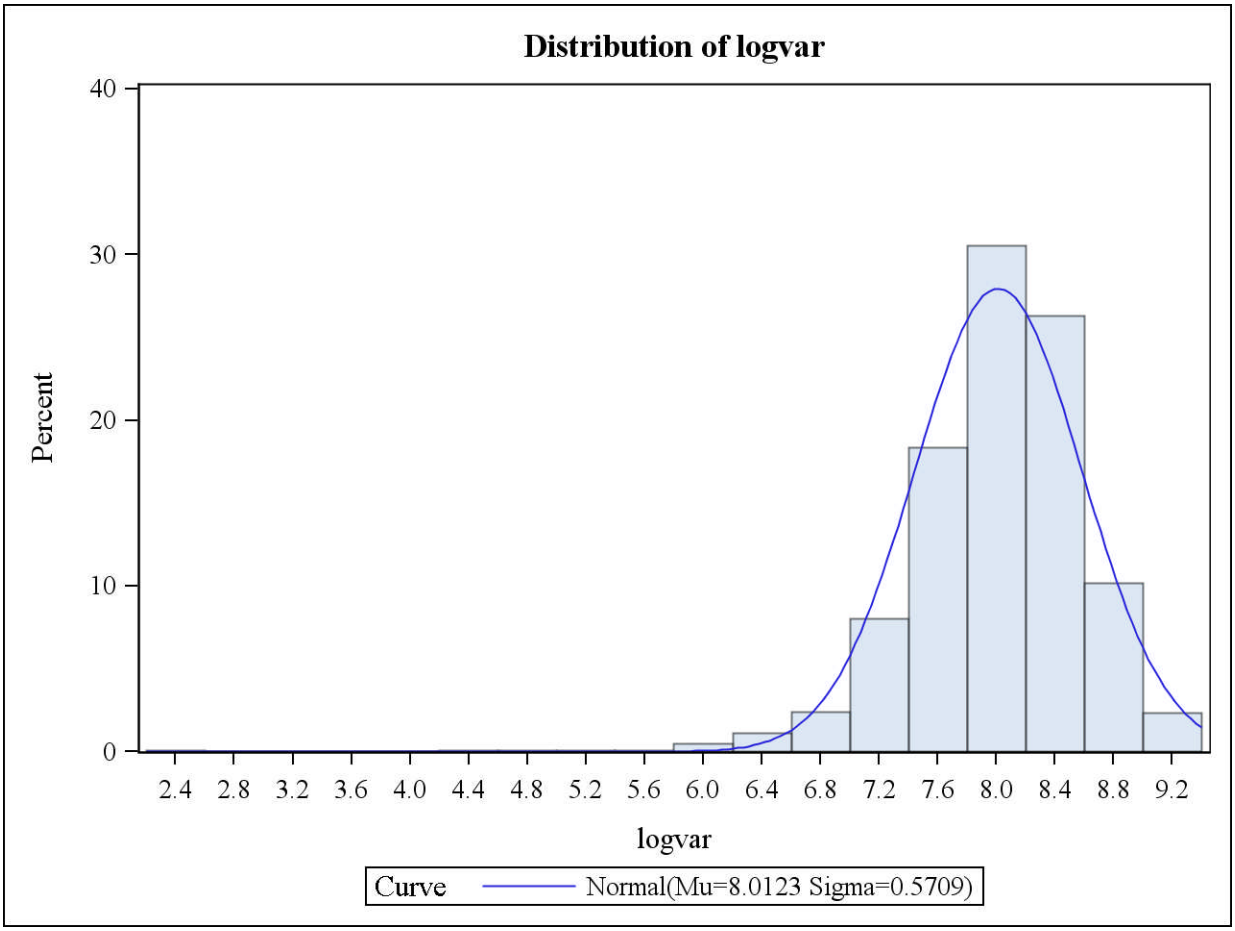
Variable: logvar

Quantiles (Definition 5)	
Quantile	**Estimate**
10%	7.32128
5%	7.05695
1%	6.34173
0% Min	2.44803

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
2.44803	1921	9.21962	1681
4.24716	647	9.23031	1456
4.55087	508	9.23681	256
4.95003	1547	9.24889	1834
5.25763	1781	9.26109	1860

The UNIVARIATE Procedure

Variable: logvar



The UNIVARIATE Procedure

Fitted Normal Distribution for logvar

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	8.012279
Std Dev	Sigma	0.57089

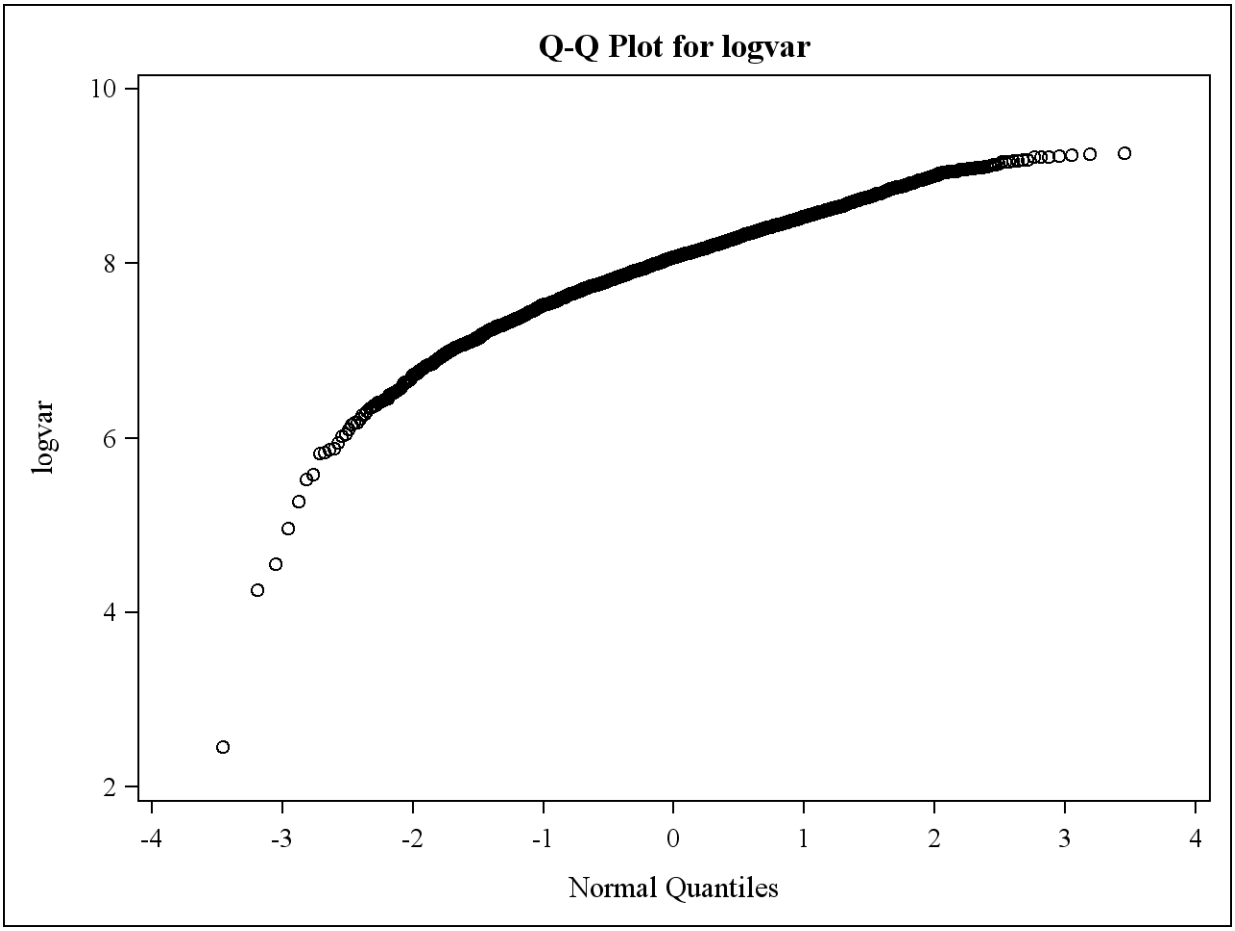
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.0541905	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	2.2498489	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	13.936183 0	**Pr > A-Sq**	<0.005

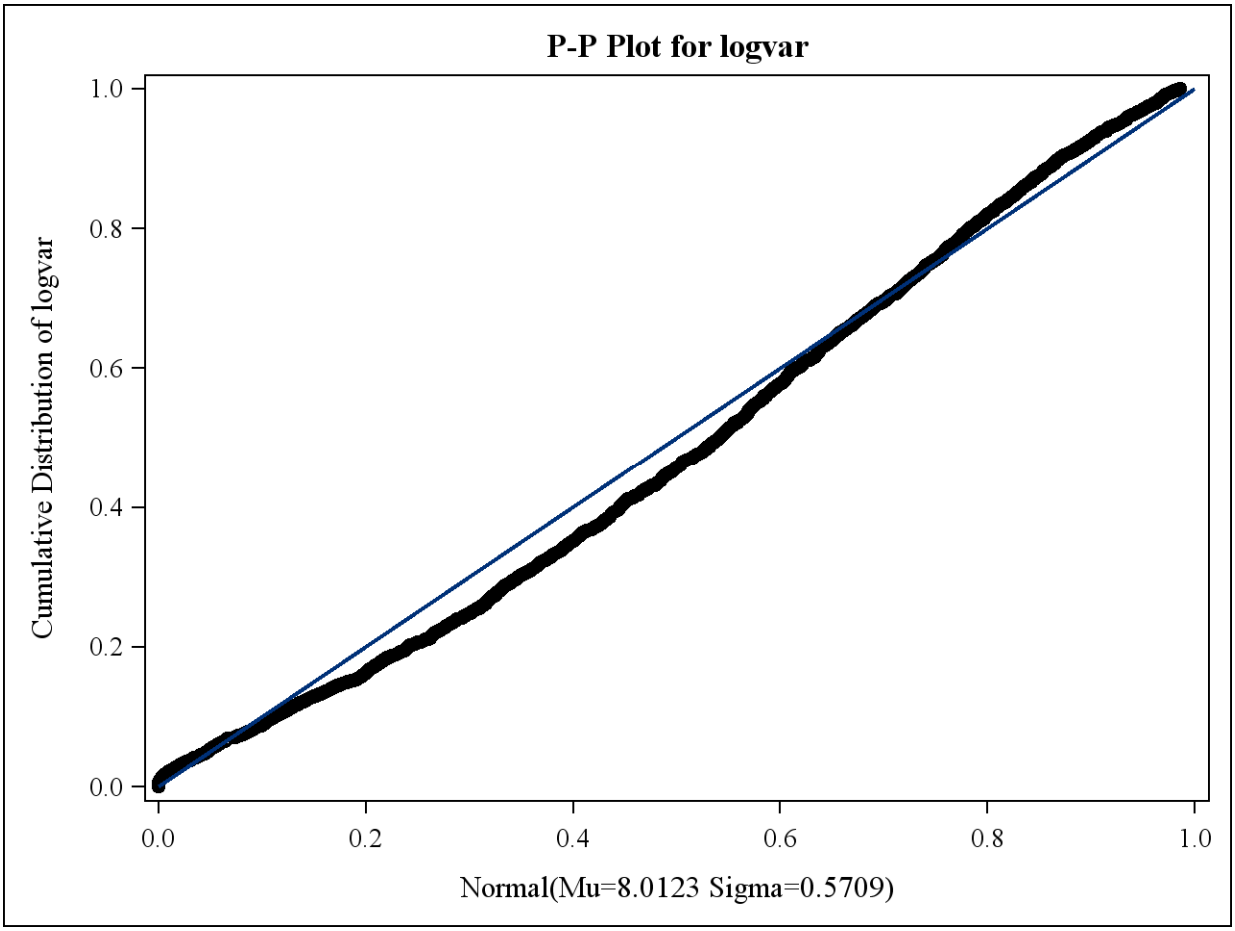
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	6.34173	6.68419
5.0	7.05695	7.07325
10.0	7.32128	7.28065

The UNIVARIATE Procedure

Fitted Normal Distribution for logvar

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
25.0	7.71713	7.62722
50.0	8.07079	8.01228
75.0	8.38700	8.39734
90.0	8.65490	8.74390
95.0	8.84109	8.95131
99.0	9.09633	9.34037


Q-Q Plot for logvar
logvar
10
8
6
4
2
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles


P-P Plot for logvar
Cumulative Distribution of logvar
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Normal(Mu=8.0123 Sigma=0.5709)

The UNIVARIATE Procedure

Variable: sqrtvar

Moments			
N	2293	**Sum Weights**	2293
Mean	57.0070884	**Sum Observations**	130717.254
Std Deviation	14.7405003	**Variance**	217.282348
Skewness	0.1923388 4	**Kurtosis**	0.2303024 9
Uncorrected SS	7949821.19	**Corrected SS**	498011.142
Coeff Variation	25.857311 2	**Std Error Mean**	0.3078294 7

Basic Statistical Measures			
Location		**Variability**	
Mean	57.00709	**Std Deviation**	14.74050
Median	56.56521	**Variance**	217.28235
Mode	44.96152	**Range**	99.16911
		Interquartile Range	18.85705

The UNIVARIATE Procedure

Variable: sqrtvar

Note: The mode displayed is the smallest of 12 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	185.1905	**Pr > |t|**	<.0001
Sign	**M**	1146.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1315036	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	102.56992
99%	94.45869
95%	83.14173
90%	75.75105
75% Q3	66.25430
50% Median	56.56521
25% Q1	47.39724
10%	38.88628

The UNIVARIATE Procedure

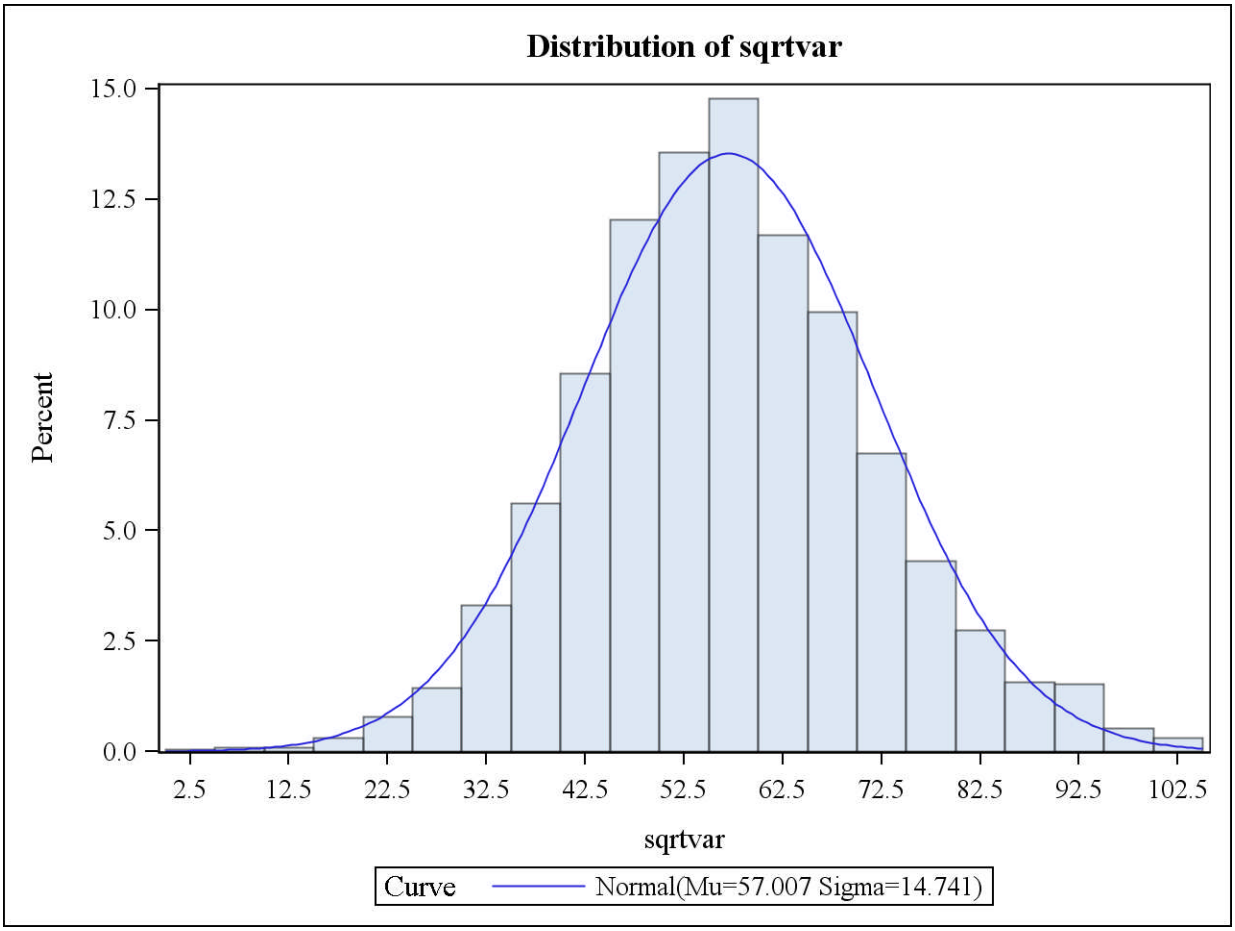
Variable: sqrtvar

Quantiles (Definition 5)	
Quantile	**Estimate**
5%	34.07191
1%	23.82809
0% Min	3.40081

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
3.40081	1921	100.465	1681
8.36101	647	101.003	1456
9.73217	508	101.332	256
11.88189	1547	101.946	1834
13.85735	1781	102.570	1860

The UNIVARIATE Procedure

Variable: sqrtvar



The UNIVARIATE Procedure

Fitted Normal Distribution for sqrtvar

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	57.00709
Std Dev	Sigma	14.7405

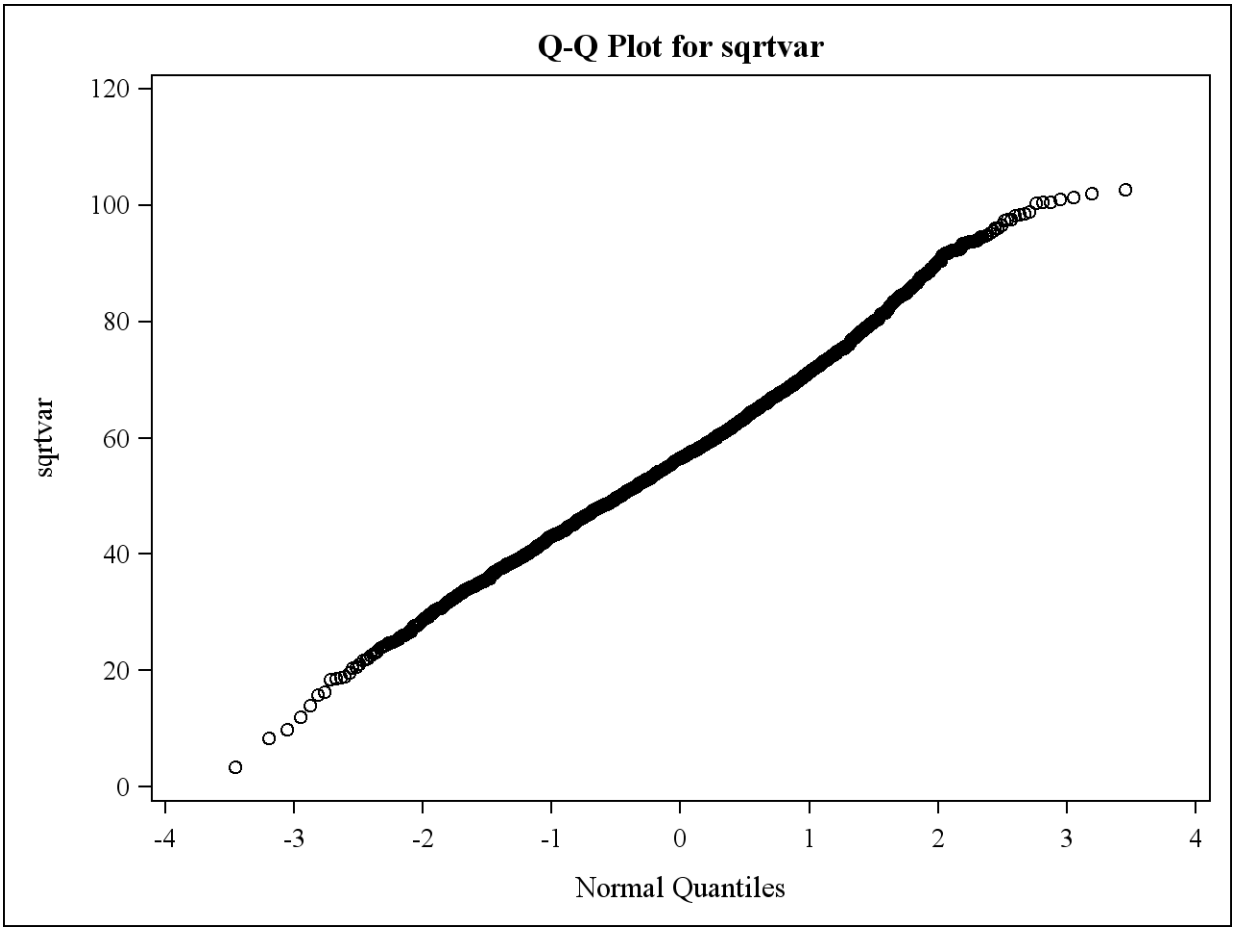
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.02780343	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	0.39285128	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	2.56219132	**Pr > A-Sq**	<0.005

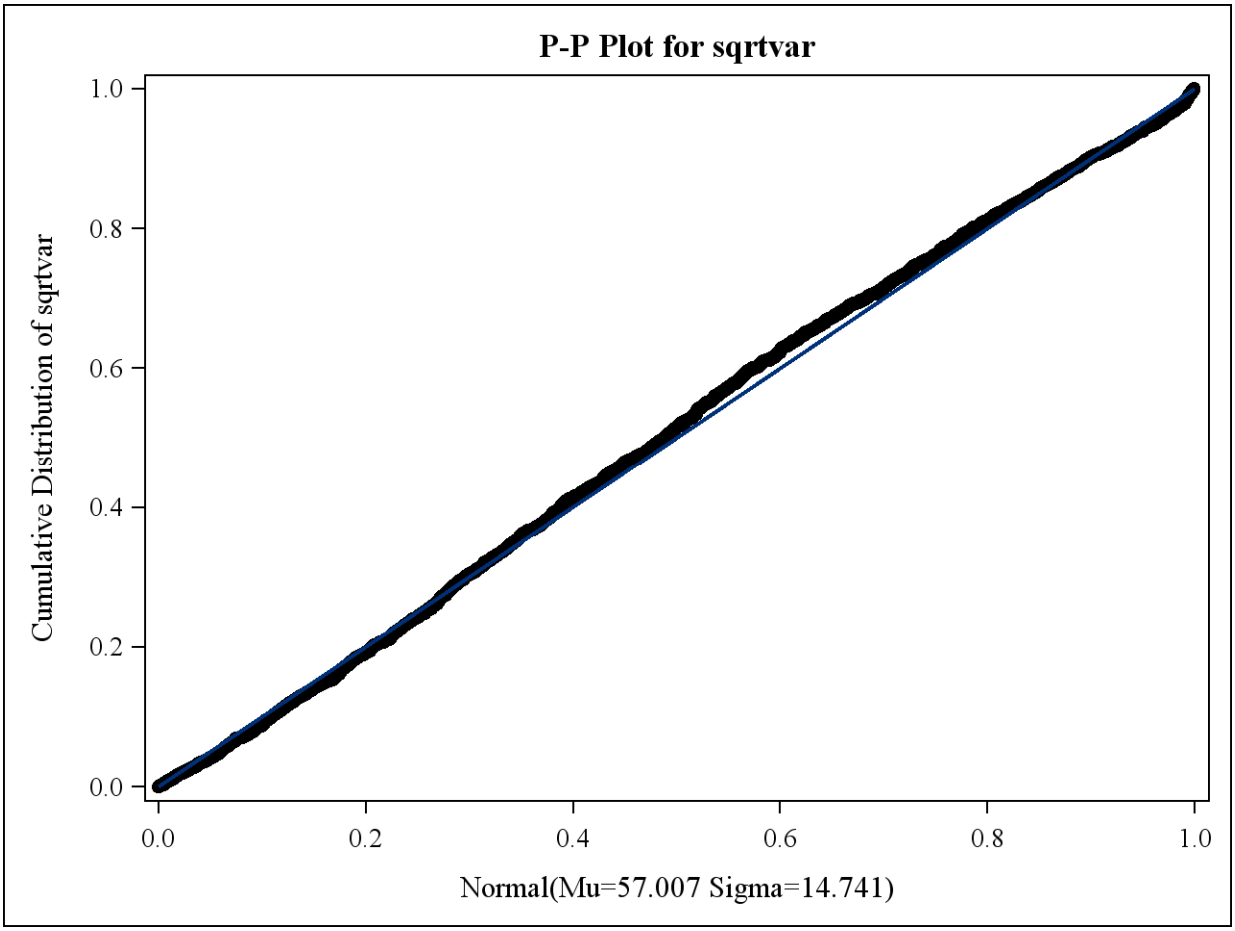
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	23.8281	22.7156
5.0	34.0719	32.7611
10.0	38.8863	38.1164

The UNIVARIATE Procedure

Fitted Normal Distribution for sqrtvar

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
25.0	47.3972	47.0648
50.0	56.5652	57.0071
75.0	66.2543	66.9494
90.0	75.7510	75.8978
95.0	83.1417	81.2531
99.0	94.4587	91.2986


Q-Q Plot for sqrtvar
sqrtvar
120
100
80
60
40
20
0
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles


P-P Plot for sqrtvar
Cumulative Distribution of sqrtvar
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Normal(Mu=57.007 Sigma=14.741)

The UNIVARIATE Procedure

Variable: fthrtvar

Moments			
N	2293	**Sum Weights**	2293
Mean	7.48369094	**Sum Observations**	17160.1033
Std Deviation	1.00094717	**Variance**	1.00189524
Skewness	-0.3371009	**Kurtosis**	0.94097165
Uncorrected SS	130717.254	**Corrected SS**	2296.34388
Coeff Variation	13.3750469	**Std Error Mean**	0.02090302

Basic Statistical Measures			
Location		**Variability**	
Mean	7.483691	**Std Deviation**	1.00095
Median	7.520985	**Variance**	1.00190
Mode	6.705335	**Range**	8.28355
		Interquartile Range	1.25511

The UNIVARIATE Procedure

Variable: fthrtvar

Note: The mode displayed is the smallest of 12 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	358.0195	**Pr > |t|**	<.0001
Sign	**M**	1146.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1315036	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	10.12768
99%	9.71899
95%	9.11821
90%	8.70351
75% Q3	8.13967
50% Median	7.52098
25% Q1	6.88457

The UNIVARIATE Procedure

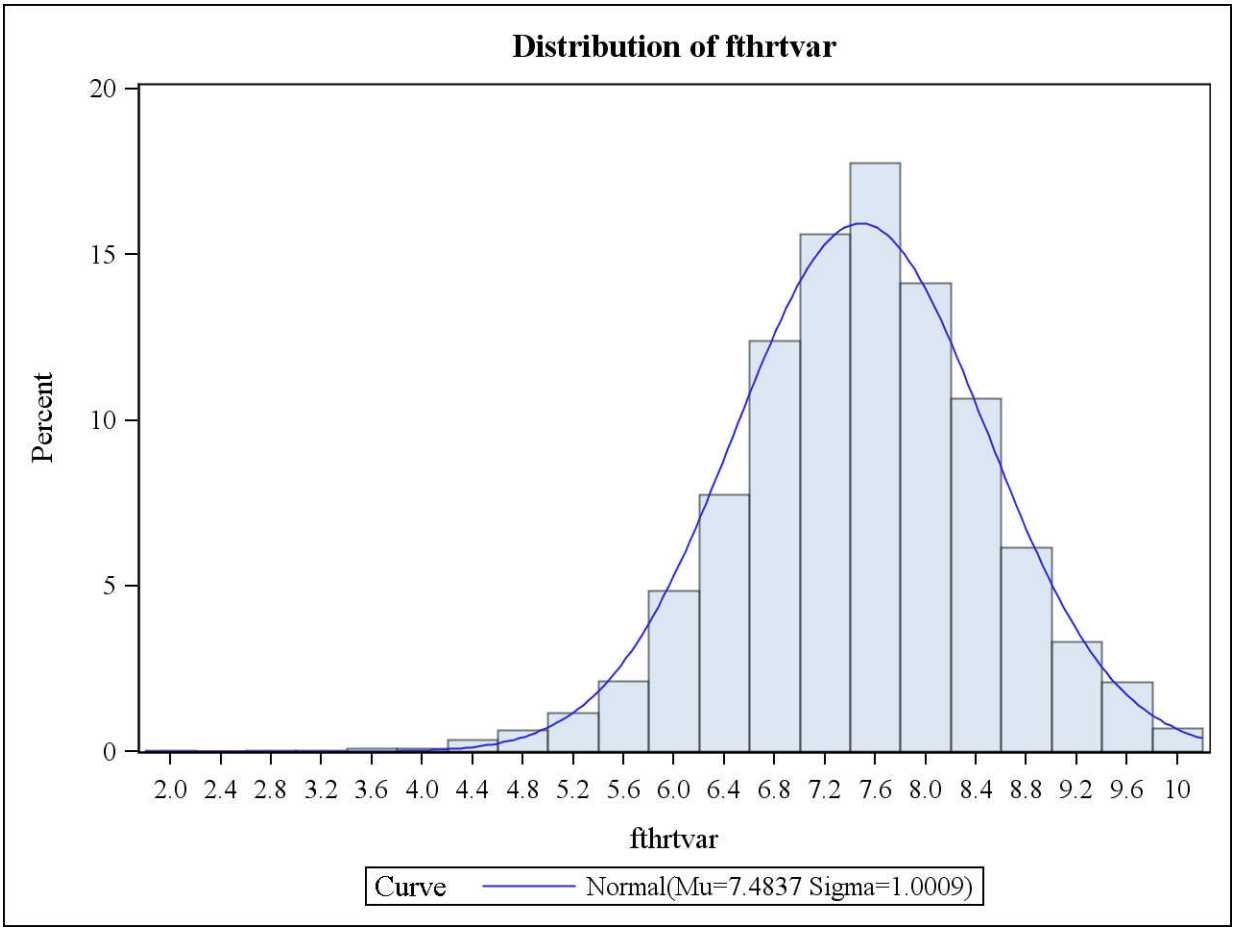
Variable: fthrtvar

Quantiles (Definition 5)	
Quantile	**Estimate**
10%	6.23589
5%	5.83711
1%	4.88140
0% Min	1.84413

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
1.84413	1921	10.0232	1681
2.89154	647	10.0500	1456
3.11964	508	10.0664	256
3.44701	1547	10.0968	1834
3.72255	1781	10.1277	1860

The UNIVARIATE Procedure

Variable: fthrtvar



The UNIVARIATE Procedure

Fitted Normal Distribution for fthrtvar

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	7.483691
Std Dev	Sigma	1.000947

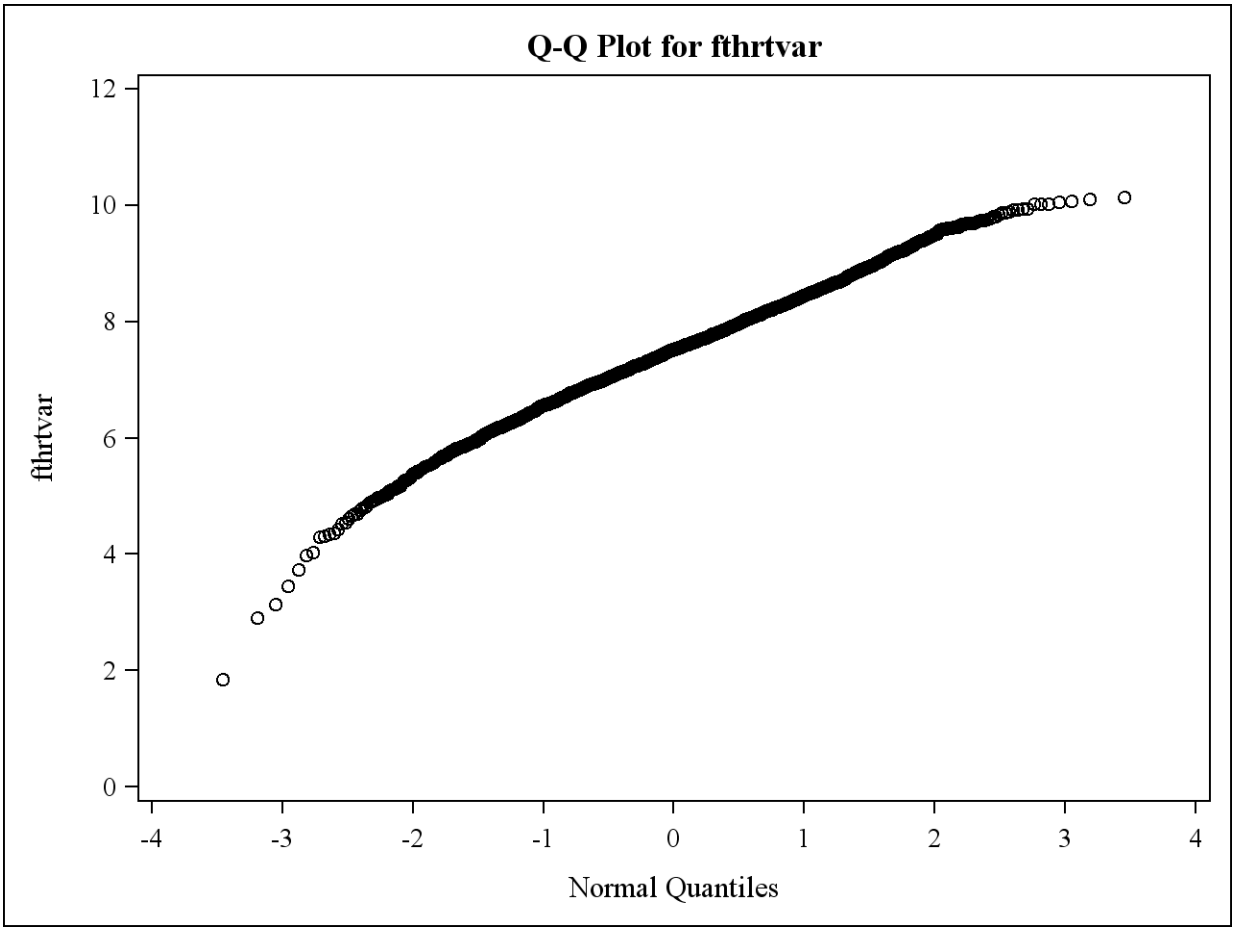
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.02607538	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	0.34187090	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	2.19455898	**Pr > A-Sq**	<0.005

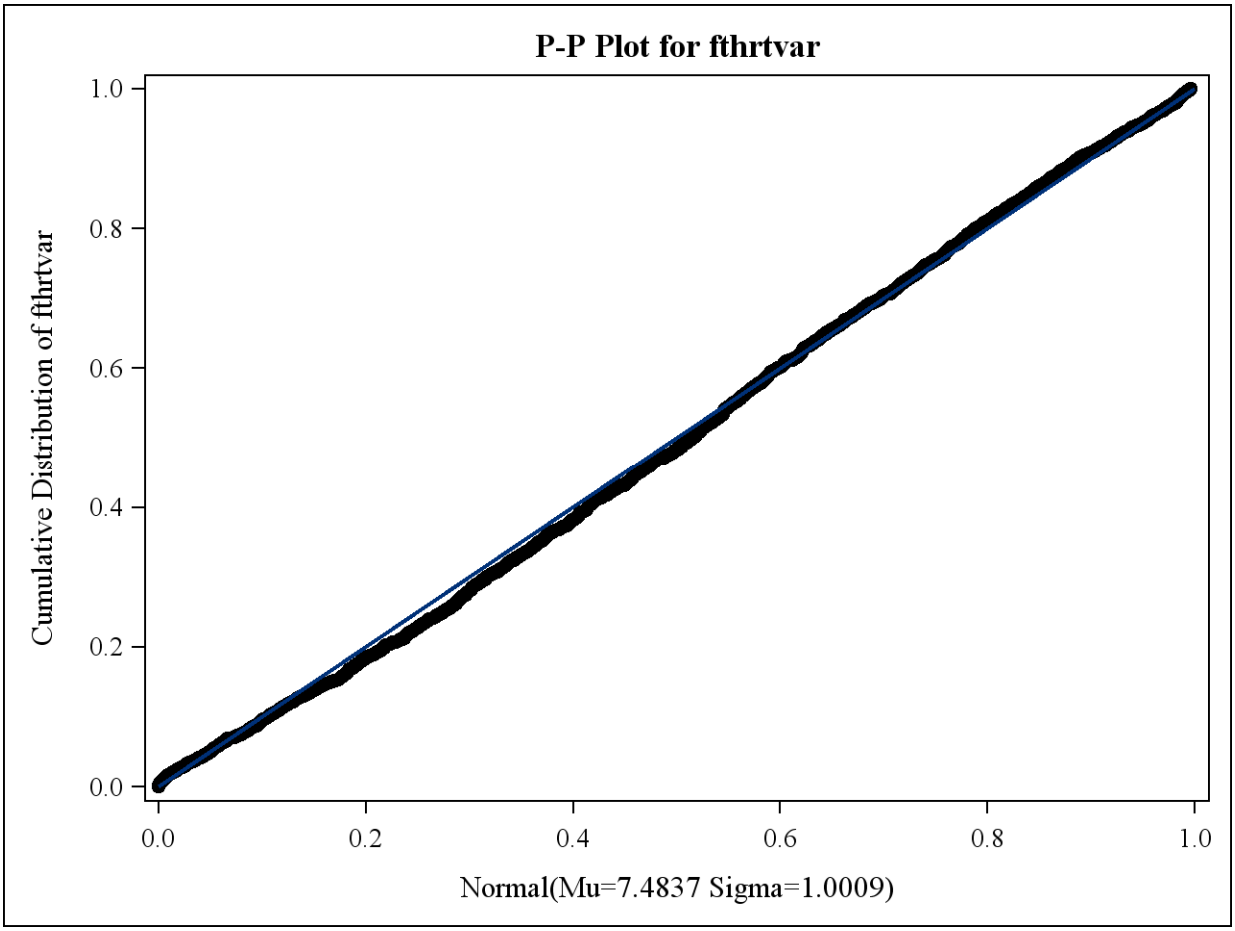
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	4.88140	5.15514
5.0	5.83711	5.83728
10.0	6.23589	6.20093

The UNIVARIATE Procedure

Fitted Normal Distribution for fthrtvar

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
25.0	6.88457	6.80856
50.0	7.52098	7.48369
75.0	8.13967	8.15882
90.0	8.70351	8.76646
95.0	9.11821	9.13010
99.0	9.71899	9.81224


Q-Q Plot for fthrtvar
fthrtvar
Normal Quantiles
12
10
8
6
4
2
0
-4
-3
-2
-1
0
1
2
3
4


P-P Plot for fthrtvar
Cumulative Distribution of fthrtvar
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Normal(Mu=7.4837 Sigma=1.0009)

The UNIVARIATE Procedure

Variable: sqvar

Moments			
N	2293	**Sum Weights**	2293
Mean	15088677.8	**Sum Observations**	3.45983E10
Std Deviation	15857533.9	**Variance**	2.51461E14
Skewness	2.39925436	**Kurtosis**	7.24807558
Uncorrected SS	1.09839E18	**Corrected SS**	5.76349E17
Coeff Variation	105.095583	**Std Error Mean**	331156.757

Basic Statistical Measures			
Location		**Variability**	
Mean	15088678	**Std Deviation**	15857534
Median	10237589	**Variance**	2.51461E14
Mode	4086618	**Range**	110682643
		Interquartile Range	14222112

Note: The mode displayed is the smallest of 12 modes with a count of 2.

The UNIVARIATE Procedure

Variable: sqvar

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	45.56355	**Pr > |t|**	<.0001
Sign	**M**	1146.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1315036	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	1.10683E+08
99%	7.96100E+07
95%	4.77833E+07
90%	3.29272E+07
75% Q3	1.92689E+07
50% Median	1.02376E+07
25% Q1	5.04676E+06
10%	2.28658E+06
5%	1.34768E+06
1%	3.22372E+05
0% Min	1.33761E+02

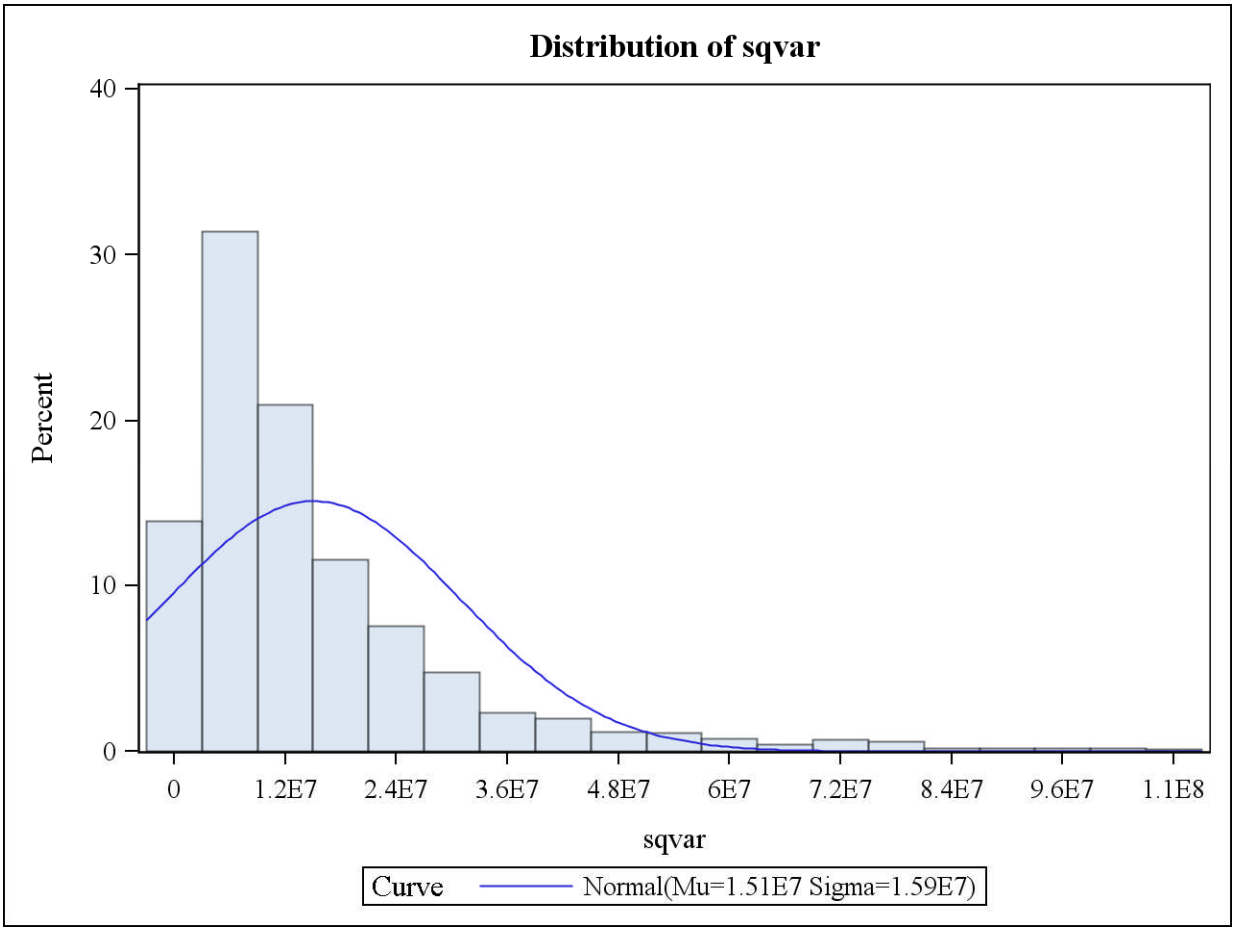
The UNIVARIATE Procedure

Variable: sqvar

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
133.761	1921	101873548	1681
4886.926	647	104073733	1456
8970.967	508	105436664	256
19931.579	1547	108013907	1834
36874.023	1781	110682777	1860

The UNIVARIATE Procedure

Variable: sqvar



The UNIVARIATE Procedure

Fitted Normal Distribution for sqvar

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	15088678
Std Dev	Sigma	15857534

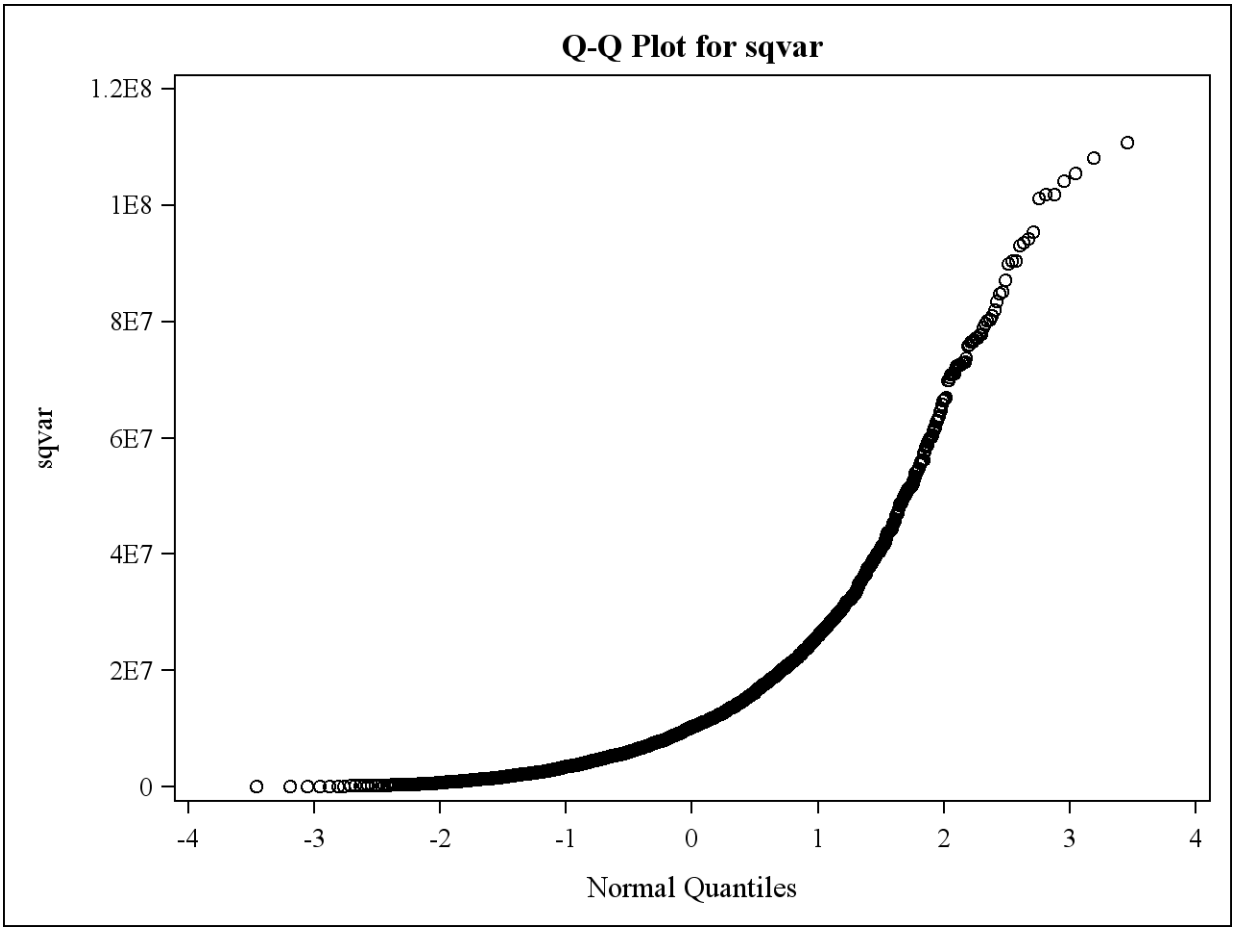
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.170674	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	26.342211	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	150.372733	**Pr > A-Sq**	<0.005

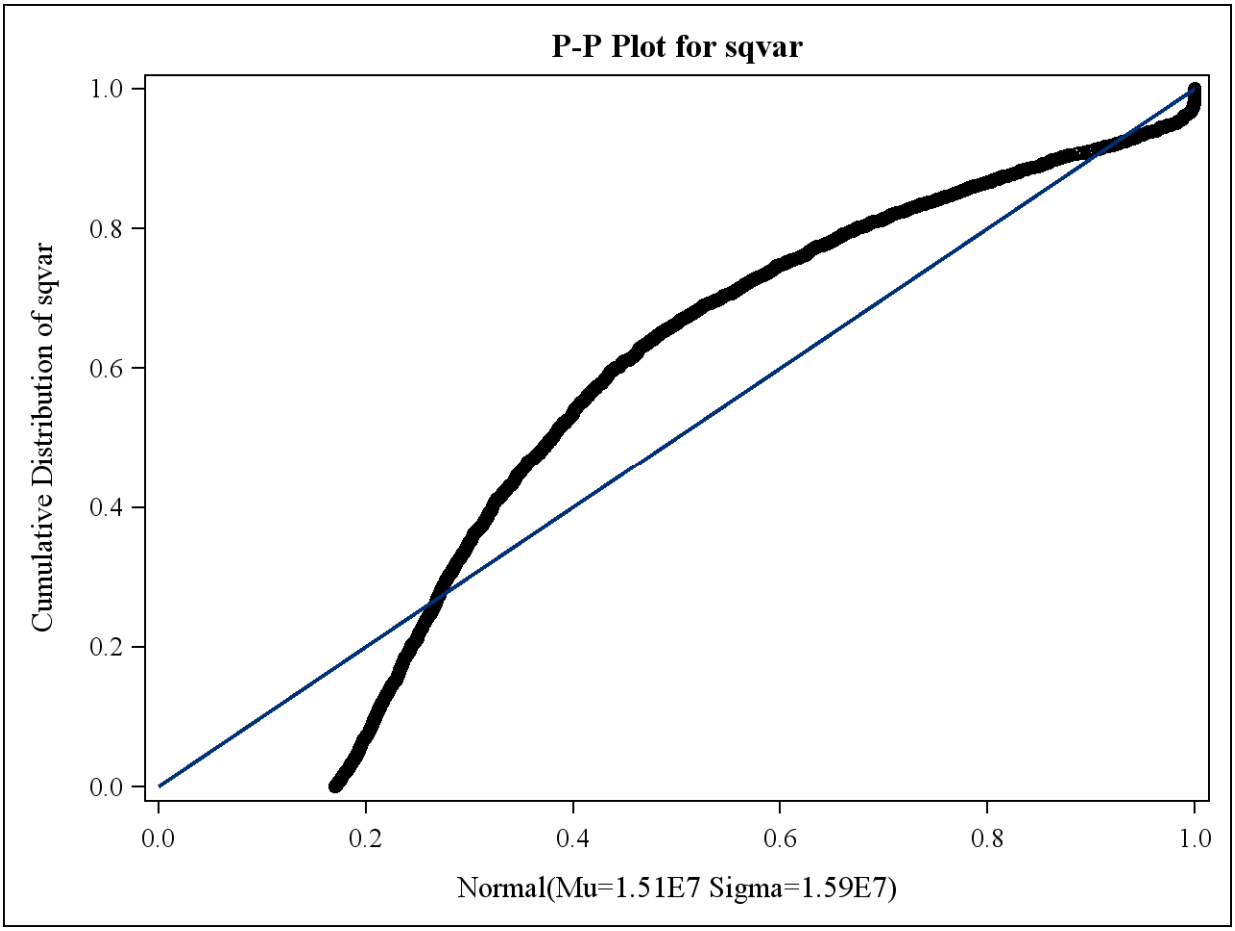
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	322372	-21801462
5.0	1347677	-10994644
10.0	2286576	-5233570
25.0	5046757	4392934
50.0	10237589	15088678
75.0	19268869	25784422

The UNIVARIATE Procedure

Fitted Normal Distribution for sqvar

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
90.0	32927182	35410925
95.0	47783310	41172000
99.0	79610026	51978818


Q-Q Plot for sqvar
1.2E8
1E8
8E7
6E7
4E7
2E7
0
sqvar
-4
-3
-2
-1
0
1
2
3
4
Normal Quantiles


P-P Plot for sqvar
Cumulative Distribution of sqvar
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Normal(Mu=1.51E7 Sigma=1.59E7)

The UNIVARIATE Procedure

Variable: EANNKWH (Annual elec kWH)

Weight: GRHFLTRR (GRHFLTRR)

Weighted Moments			
N	1776	**Sum Weights**	14018680
Mean	3943.3159	**Sum Observations**	5.52801E10
Std Deviation	225902.822	**Variance**	5.10321E10
Skewness	6.65957895	**Kurtosis**	97.4741341
Uncorrected SS	3.08569E14	**Corrected SS**	9.0582E13
Coeff Variation	5728.75284	**Std Error Mean**	60.3348302

Weighted Basic Statistical Measures			
Location		**Variability**	
Mean	3943.316	**Std Deviation**	225903
Median	3425.165	**Variance**	5.10321E10
Mode	1277.500	**Range**	41516
		Interquartile Range	2232

Note: The mode displayed is the smallest of 6 modes with a count of 2.

The UNIVARIATE Procedure

Variable: EANNKWH (Annual elec kWH)

Weight: GRHFLTRR (GRHFLTRR)

Weighted Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	65.35721	**Pr > |t|**	<.0001

Weighted Quantiles	
Quantile	**Estimate**
100% Max	41527.2734
99%	11557.5669
95%	8286.9077
90%	6479.4662
75% Q3	4760.5213
50% Median	3425.1646
25% Q1	2528.6834
10%	1831.2075
5%	1416.8819
1%	794.4517
0% Min	11.5655

The UNIVARIATE Procedure

Variable: EANNKWH (Annual elec kWH)

Weight: GRHFLTRR (GRHFLTRR)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
11.5655	2011	21612.0	1424
94.7152	529	24488.5	1812
141.1792	1626	27058.0	1064
343.4409	1862	33405.1	807
358.3313	74	41527.3	1962

The UNIVARIATE Procedure

Variable: logvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Moments			
N	1776	**Sum Weights**	14018680
Mean	8.13430247	**Sum Observations**	114032183
Std Deviation	49.5753751	**Variance**	2457.71781
Skewness	-1.0705075	**Kurtosis**	17.0104572
Uncorrected SS	931934719	**Corrected SS**	4362449.12
Coeff Variation	609.460679	**Std Error Mean**	0.01324075

Weighted Basic Statistical Measures			
Location		**Variability**	
Mean	8.134302	**Std Deviation**	49.57538
Median	8.138905	**Variance**	2458
Mode	7.152660	**Range**	8.18608
		Interquartile Range	0.63266

The UNIVARIATE Procedure

Variable: logvar

Weight: GRHFLTRR (GRHFLTRR)

Note: The mode displayed is the smallest of 6 modes with a count of 2.

Weighted Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	614.3387	**Pr > |t|**	<.0001

Weighted Quantiles	
Quantile	**Estimate**
100% Max	10.63411
99%	9.35510
95%	9.02243
90%	8.77639
75% Q3	8.46811
50% Median	8.13890
25% Q1	7.83545
10%	7.51273
5%	7.25621

The UNIVARIATE Procedure

Variable: logvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Quantiles	
Quantile	**Estimate**
1%	6.67765
0% Min	2.44803

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
2.44803	2011	9.9810	1424
4.55087	529	10.1060	1812
4.95003	1626	10.2057	1064
5.83902	1862	10.4165	807
5.88146	74	10.6341	1962

The UNIVARIATE Procedure

Variable: sqrtvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Moments			
N	1776	**Sum Weights**	14018680
Mean	60.5666937	**Sum Observations**	849065098
Std Deviation	1473.71679	**Variance**	2171841.17
Skewness	2.03704353	**Kurtosis**	15.5036472
Uncorrected SS	5.52801E10	**Corrected SS**	3855018084
Coeff Variation	2433.21321	**Std Error Mean**	0.39360488

Weighted Basic Statistical Measures			
Location		**Variability**	
Mean	60.56669	**Std Deviation**	1474
Median	58.52491	**Variance**	2171841
Mode	35.74213	**Range**	200.38160
		Interquartile Range	18.71051

Note: The mode displayed is the smallest of 6 modes with a count of 2.

The UNIVARIATE Procedure

Variable: sqrtvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	153.8769	**Pr > |t|**	<.0001

Weighted Quantiles	
Quantile	**Estimate**
100% Max	203.78242
99%	107.50613
95%	91.03245
90%	80.49513
75% Q3	68.99653
50% Median	58.52491
25% Q1	50.28602
10%	42.79261
5%	37.64149
1%	28.18602
0% Min	3.40081

The UNIVARIATE Procedure

Variable: sqrtvar

Weight: GRHFLTRR (GRHFLTRR)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
3.40081	2011	147.010	1424
9.73217	529	156.488	1812
11.88189	1626	164.493	1064
18.53216	1862	182.771	807
18.92964	74	203.782	1962

The UNIVARIATE Procedure

Variable: fthrtvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Moments			
N	1776	**Sum Weights**	14018680
Mean	7.71307748	**Sum Observations**	108127165
Std Deviation	92.147756	**Variance**	8491.20893
Skewness	0.73171873	**Kurtosis**	7.63985372
Uncorrected SS	849065098	**Corrected SS**	15071895.8
Coeff Variation	1194.69506	**Std Error Mean**	0.02461111

Weighted Basic Statistical Measures			
Location		**Variability**	
Mean	7.713077	**Std Deviation**	92.14776
Median	7.650157	**Variance**	8491
Mode	5.978472	**Range**	12.43111
		Interquartile Range	1.21515

Note: The mode displayed is the smallest of 6 modes with a count of 2.

The UNIVARIATE Procedure

Variable: fthrtvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	313.3982	**Pr > |t|**	<.0001

Weighted Quantiles	
Quantile	**Estimate**
100% Max	14.27524
99%	10.36852
95%	9.54109
90%	8.97191
75% Q3	8.30642
50% Median	7.65016
25% Q1	7.09126
10%	6.54161
5%	6.13527
1%	5.30905
0% Min	1.84413

The UNIVARIATE Procedure

Variable: fthrtvar

Weight: GRHFLTRR (GRHFLTRR)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.84413	2011	12.1248	1424
3.11964	529	12.5095	1812
3.44701	1626	12.8255	1064
4.30490	1862	13.5193	807
4.35082	74	14.2752	1962

The UNIVARIATE Procedure

Variable: EANNKWH (Annual elec kWH)

Weight: GRHFLTRR (GRHFLTRR)

Weighted Moments			
N	1701	**Sum Weights**	13367103
Mean	3732.7172	**Sum Observations**	4.98956E10
Std Deviation	158806.303	**Variance**	2.52194E10
Skewness	1.72102735	**Kurtosis**	7.90428739
Uncorrected SS	2.29119E14	**Corrected SS**	4.28731E13
Coeff Variation	4254.4424	**Std Error Mean**	43.4359265

Weighted Basic Statistical Measures			
Location		**Variability**	
Mean	3732.717	**Std Deviation**	158806
Median	3361.661	**Variance**	2.52194E10
Mode	2111.786	**Range**	10509
		Interquartile Range	2092

Note: The mode displayed is the smallest of 5 modes with a count of 2.

The UNIVARIATE Procedure

Variable: EANNKWH (Annual elec kWH)

Weight: GRHFLTRR (GRHFLTRR)

Weighted Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	85.93617	**Pr > |t|**	<.0001

Weighted Quantiles	
Quantile	**Estimate**
100% Max	10520.5882
99%	8953.0288
95%	7387.4654
90%	6154.7937
75% Q3	4605.5425
50% Median	3361.6614
25% Q1	2513.8874
10%	1809.4360
5%	1411.8681
1%	778.4360
0% Min	11.5655

The UNIVARIATE Procedure

Variable: EANNKWH (Annual elec kWH)

Weight: GRHFLTRR (GRHFLTRR)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
11.5655	1921	10088.9	43
94.7152	508	10093.2	1681
141.1792	1547	10201.7	1456
343.4409	1776	10268.2	256
358.3313	72	10520.6	1860

The UNIVARIATE Procedure

Variable: logvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Moments			
N	1701	**Sum Weights**	13367103
Mean	8.10436158	**Sum Observations**	108331836
Std Deviation	47.1077286	**Variance**	2219.1381
Skewness	-1.6263393	**Kurtosis**	19.8182287
Uncorrected SS	881732904	**Corrected SS**	3772534.76
Coeff Variation	581.263905	**Std Error Mean**	0.01288468

Weighted Basic Statistical Measures			
Location		**Variability**	
Mean	8.104362	**Std Deviation**	47.10773
Median	8.120191	**Variance**	2219
Mode	7.655289	**Range**	6.81306
		Interquartile Range	0.60543

Note: The mode displayed is the smallest of 5 modes with a count of 2.

The UNIVARIATE Procedure

Variable: logvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	628.9922	**Pr > |t|**	<.0001

Weighted Quantiles	
Quantile	**Estimate**
100% Max	9.26109
99%	9.09975
95%	8.90754
90%	8.72499
75% Q3	8.43502
50% Median	8.12019
25% Q1	7.82959
10%	7.50077
5%	7.25267
1%	6.65729
0% Min	2.44803

The UNIVARIATE Procedure

Variable: logvar

Weight: GRHFLTRR (GRHFLTRR)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2.44803	1921	9.21919	43
4.55087	508	9.21962	1681
4.95003	1547	9.23031	1456
5.83902	1776	9.23681	256
5.88146	72	9.26109	1860

The UNIVARIATE Procedure

Variable: sqrtvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Moments			
N	1701	**Sum Weights**	13367103
Mean	59.3866213	**Sum Observations**	793827084
Std Deviation	1272.53958	**Variance**	1619356.98
Skewness	0.72928559	**Kurtosis**	4.4086737
Uncorrected SS	4.98956E10	**Corrected SS**	2752906859
Coeff Variation	2142.80514	**Std Error Mean**	0.34805883

Weighted Basic Statistical Measures			
Location		**Variability**	
Mean	59.38662	**Std Deviation**	1273
Median	57.97984	**Variance**	1619357
Mode	45.95417	**Range**	99.16911
		Interquartile Range	17.72547

Note: The mode displayed is the smallest of 5 modes with a count of 2.

The UNIVARIATE Procedure

Variable: sqrtvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	170.6224	**Pr > |t|**	<.0001

Weighted Quantiles	
Quantile	**Estimate**
100% Max	102.56992
99%	94.62045
95%	85.95037
90%	78.45249
75% Q3	67.86415
50% Median	57.97984
25% Q1	50.13868
10%	42.53747
5%	37.57483
1%	27.90047
0% Min	3.40081

The UNIVARIATE Procedure

Variable: sqrtvar

Weight: GRHFLTRR (GRHFLTRR)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
3.40081	1921	100.444	43
9.73217	508	100.465	1681
11.88189	1547	101.003	1456
18.53216	1776	101.332	256
18.92964	72	102.570	1860

The UNIVARIATE Procedure

Variable: fthrtvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Moments			
N	1701	**Sum Weights**	13367103
Mean	7.64766818	**Sum Observations**	102227168
Std Deviation	84.1134479	**Variance**	7075.07212
Skewness	0.0165286	**Kurtosis**	4.96096993
Uncorrected SS	793827084	**Corrected SS**	12027622.6
Coeff Variation	1099.85745	**Std Error Mean**	0.0230063

Weighted Basic Statistical Measures			
Location		**Variability**	
Mean	7.647668	**Std Deviation**	84.11345
Median	7.614449	**Variance**	7075
Mode	6.778950	**Range**	8.28355
		Interquartile Range	1.15710

Note: The mode displayed is the smallest of 5 modes with a count of 2.

The UNIVARIATE Procedure

Variable: fthrtvar

Weight: GRHFLTRR (GRHFLTRR)

Weighted Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	332.4163	**Pr > |t|**	<.0001

Weighted Quantiles	
Quantile	**Estimate**
100% Max	10.12768
99%	9.72730
95%	9.27094
90%	8.85734
75% Q3	8.23797
50% Median	7.61445
25% Q1	7.08087
10%	6.52208
5%	6.12983
1%	5.28209
0% Min	1.84413

The UNIVARIATE Procedure

Variable: fthrtvar

Weight: GRHFLTRR (GRHFLTRR)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.84413	1921	10.0222	43
3.11964	508	10.0232	1681
3.44701	1547	10.0500	1456
4.30490	1776	10.0664	256
4.35082	72	10.1277	1860

Obs	Number	Mean	Skewness	Kurtosis	name	seskew	sekurt	zskew	zkurt
1	2399	3720.85	5.71968	67.2927	transNOoutliersIN	0.049979	0.09992	114.441	673.487
2	2399	8.05	-0.79211	5.4327	transLOGoutliersIN	0.049979	0.09992	-15.849	54.372
3	2399	58.43	1.53810	8.5455	transSQRToutliersIN	0.049979	0.09992	30.775	85.526
4	2399	7.56	0.39648	3.0653	transFTHRToutliersIN	0.049979	0.09992	7.933	30.678
5	2293	3467.00	0.98858	1.2087	transNOoutliersOUT	0.051120	0.10220	19.338	11.827
6	2293	8.01	-1.26769	6.2487	transLOGoutliersOUT	0.051120	0.10220	-24.798	61.144
7	2293	57.01	0.19234	0.2303	transSQRToutliersOUT	0.051120	0.10220	3.763	2.254
8	2293	7.48	-0.33710	0.9410	transFTHRToutliersOUT	0.051120	0.10220	-6.594	9.208
9	1776	3943.32	6.65958	97.4741	transNOoutliersIN_w	0.058075	0.11608	114.672	839.682
10	1776	8.13	-1.07051	17.0105	transLOGoutliersIN_w	0.058075	0.11608	-18.433	146.535
11	1776	60.57	2.03704	15.5036	transSQRToutliersIN_w	0.058075	0.11608	35.076	133.555
12	1776	7.71	0.73172	7.6399	transFTHRToutliersIN_w	0.058075	0.11608	12.600	65.813
13	1701	3732.72	1.72103	7.9043	transNOoutliersOUT_w	0.059339	0.11861	29.003	66.642
14	1701	8.10	-1.62634	19.8182	transLOGoutliersOUT_w	0.059339	0.11861	-27.408	167.089

Obs	Number	Mean	Skewness	Kurtosis	name	seskew	sekurt	zskew	zkurt
15	1701	59.39	0.72929	4.4087	transSQRToutliersOUT_w	0.059339	0.11861	12.290	37.170
16	1701	7.65	0.01653	4.9610	transFTHRToutliersOUT_w	0.059339	0.11861	0.279	41.826

First Run – decennial model (7.2.1)

Number of Observations Read	2290
Number of Observations Used	2290

Descriptive Statistics						
Variable	**Sum**	**Mean**	**Uncorrected SS**	**Variance**	**Standard Deviation**	**Label**
Intercept	2290.00000	1.00000	2290.00000	0	0	Intercept
rooms_hsize96	29804	13.01485	540512	66.67431	8.16543	Interaction term of the number of rooms and number of occupants (1996)
sqrt_eannkwh	130787	57.11229	7960170	214.33611	14.64022	Square root transformed annual domestic electricity consumption (1996)

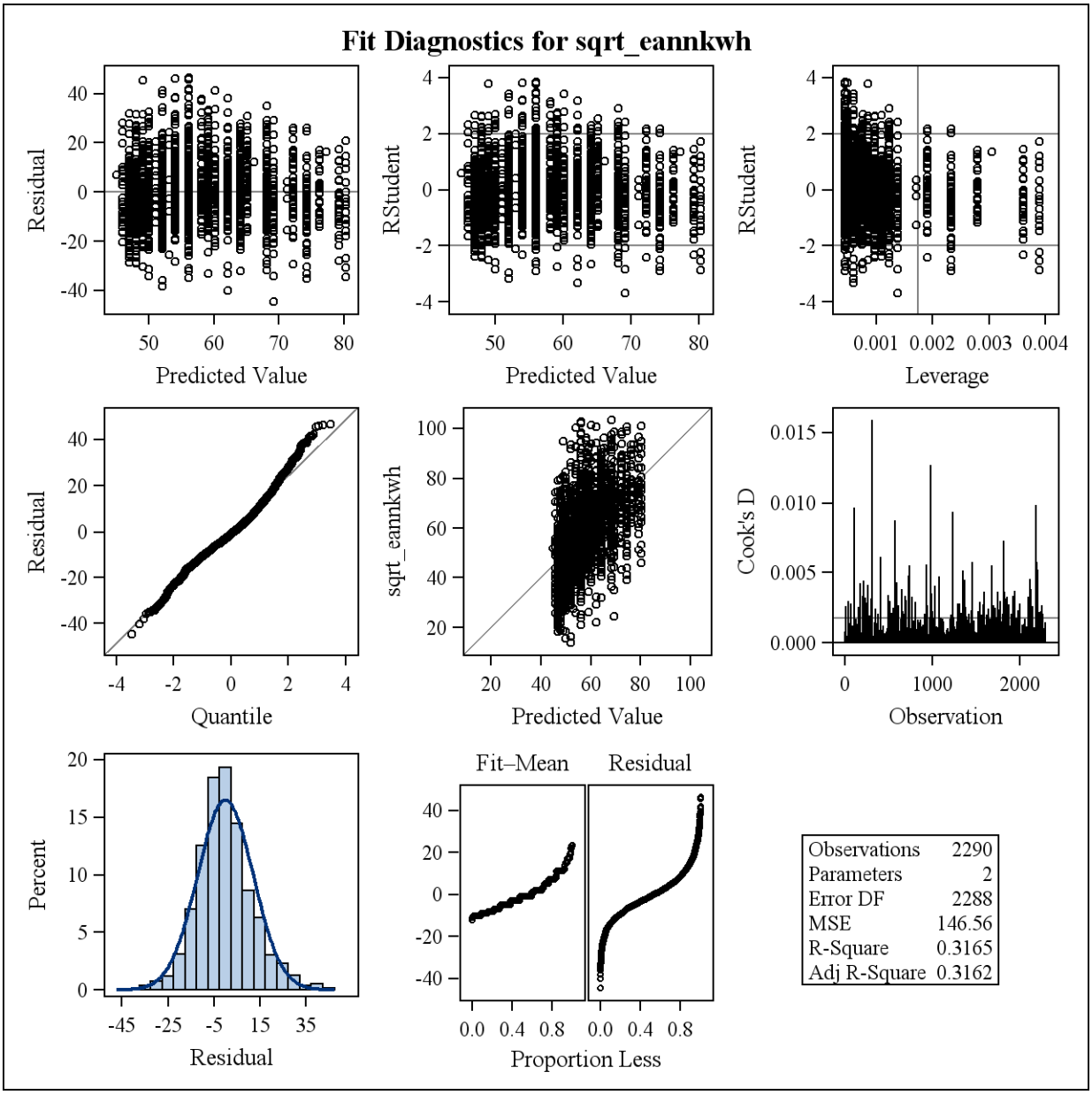
Correlation			
Variable	**Label**	**rooms_hsize96**	**sqrt_eannkwh**
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	1.0000	0.5626
sqrt_eannkwh	Square root transformed annual domestic electricity consumption (1996)	0.5626	1.0000

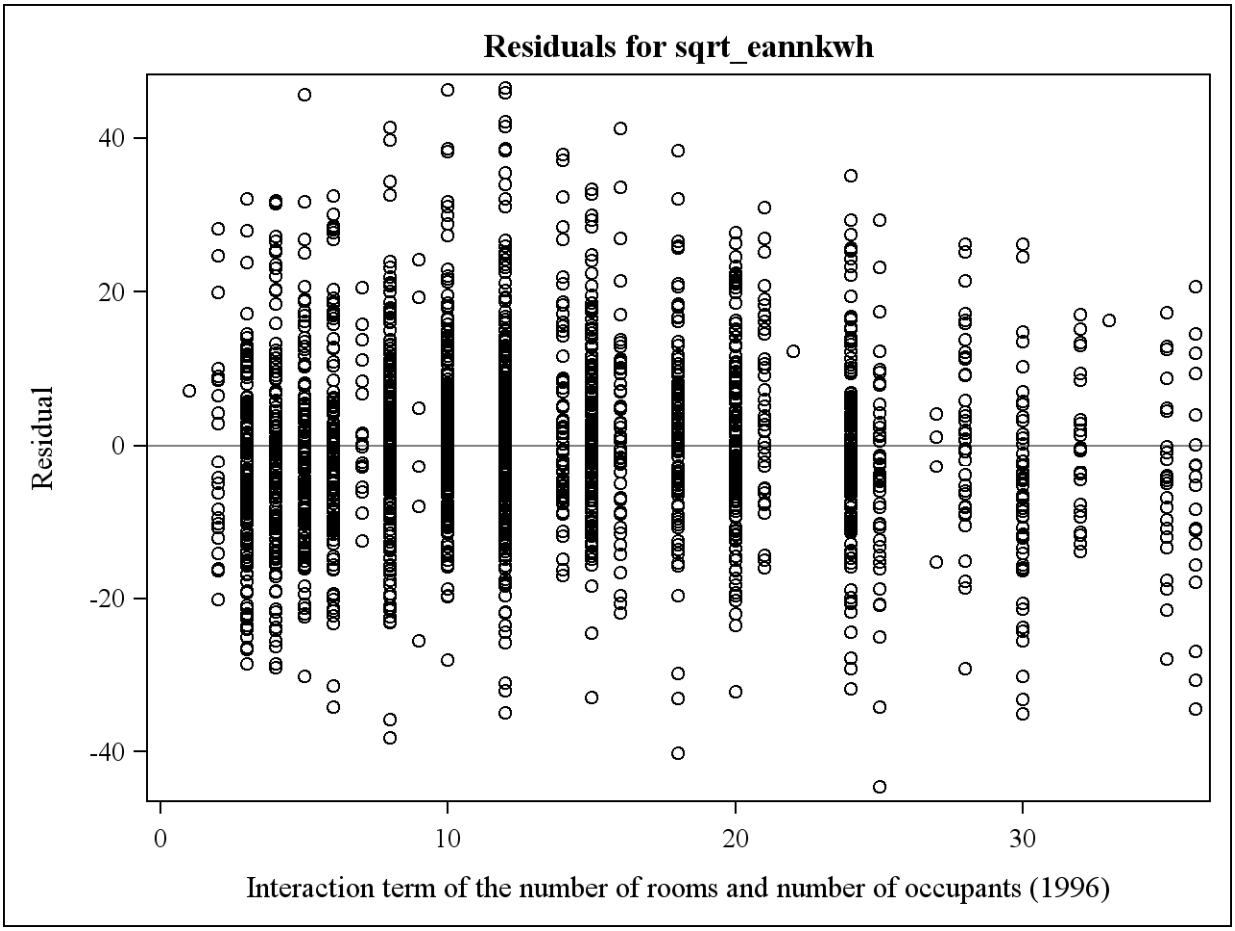
Number of Observations Read	2290
Number of Observations Used	2290

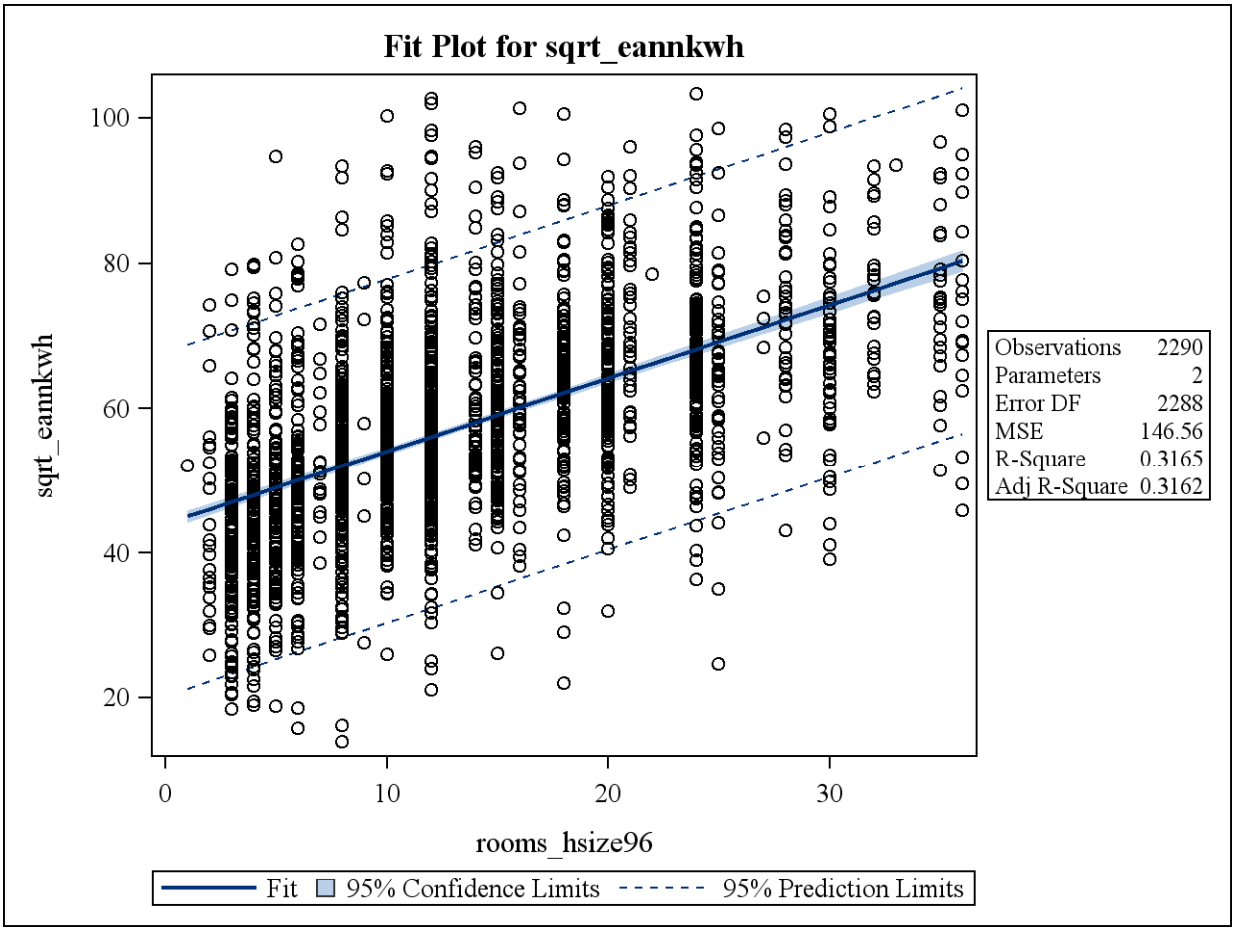
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	155292	155292	1059.59	<.0001
Error	2288	335324	146.55761		
Corrected Total	2289	490615			

Root MSE	12.10610	**R-Square**	0.3165
Dependent Mean	57.11229	**Adj R-Sq**	0.3162
Coeff Var	21.19701		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	43.98392	0.47609	92.39	<.0001
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	1	1.00872	0.03099	32.55	<.0001


Fit Diagnostics for sqrt_eannkwh
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
sqrt_eannkwh
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 2290
Parameters 2
Error DF 2288
MSE 146.56
R-Square 0.3165
Adj R-Square 0.3162


Residuals for sqrt_eannkwh
Residual
40
20
0
-20
-40
0
10
20
30
Interaction term of the number of rooms and number of occupants (1996)


Fit Plot for sqrt_eannkwh
sqrt_eannkwh
100
80
60
40
20
0
10
20
30
rooms_hsize96
Observations 2290
Parameters 2
Error DF 2288
MSE 146.56
R-Square 0.3165
Adj R-Square 0.3162
Fit
95% Confidence Limits
95% Prediction Limits

Second Run – decennial model (7.2.2)

Number of Observations Read	2156
Number of Observations Used	2156

Descriptive Statistics						
Variable	**Sum**	**Mean**	**Uncorrected SS**	**Variance**	**Standard Deviation**	**Label**
Intercept	2156.00000	1.00000	2156.00000	0	0	Intercept
rooms_hsize96	27887	12.93460	502999	66.02867	8.12580	Interaction term of the number of rooms and number of occupants (1996)
sqrt_eannkwh	121465	56.33835	7206283	168.50052	12.98077	

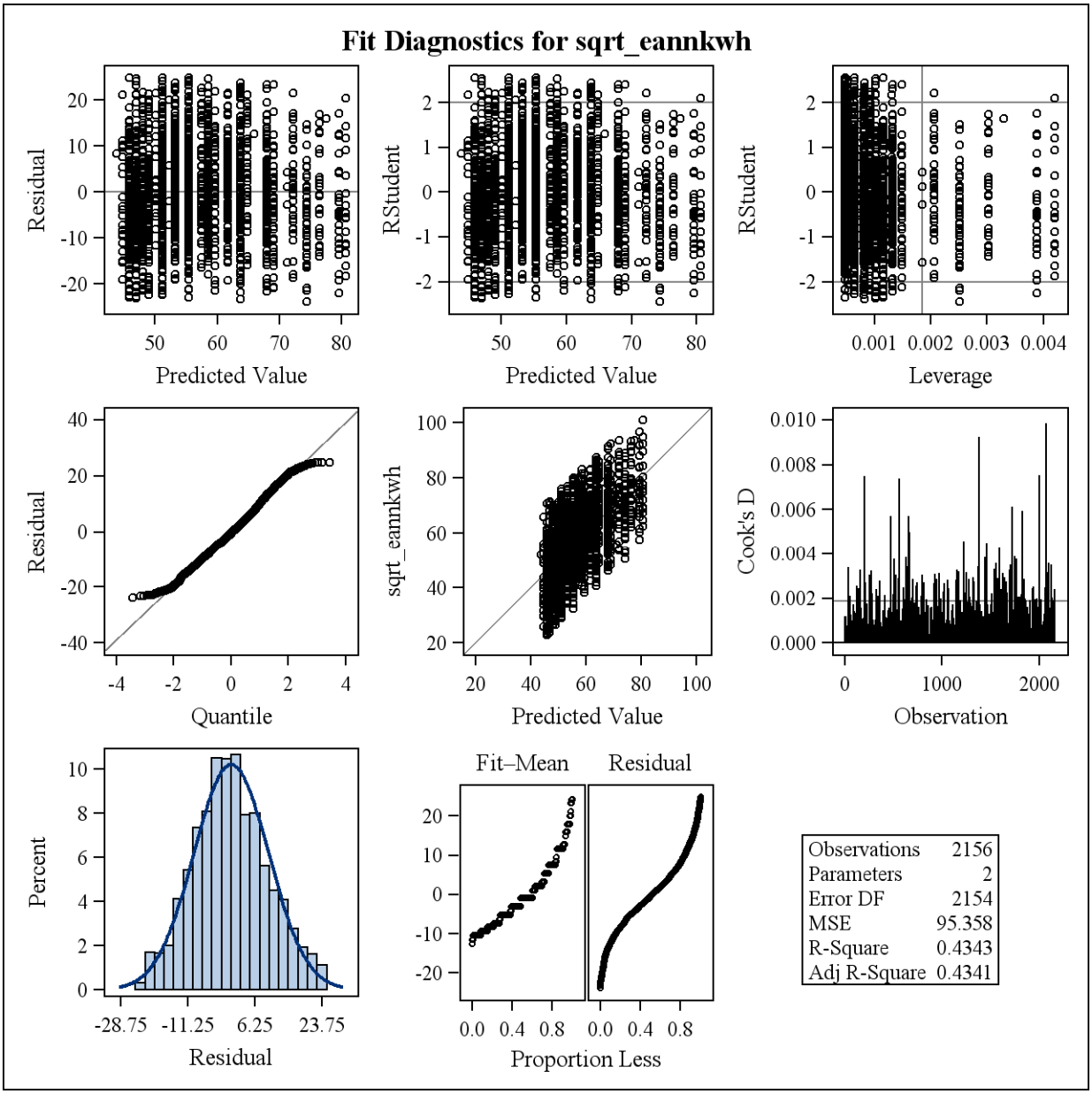
Correlation			
Variable	**Label**	**rooms_hsize96**	**sqrt_eannkwh**
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	1.0000	0.6590
sqrt_eannkwh		0.6590	1.0000

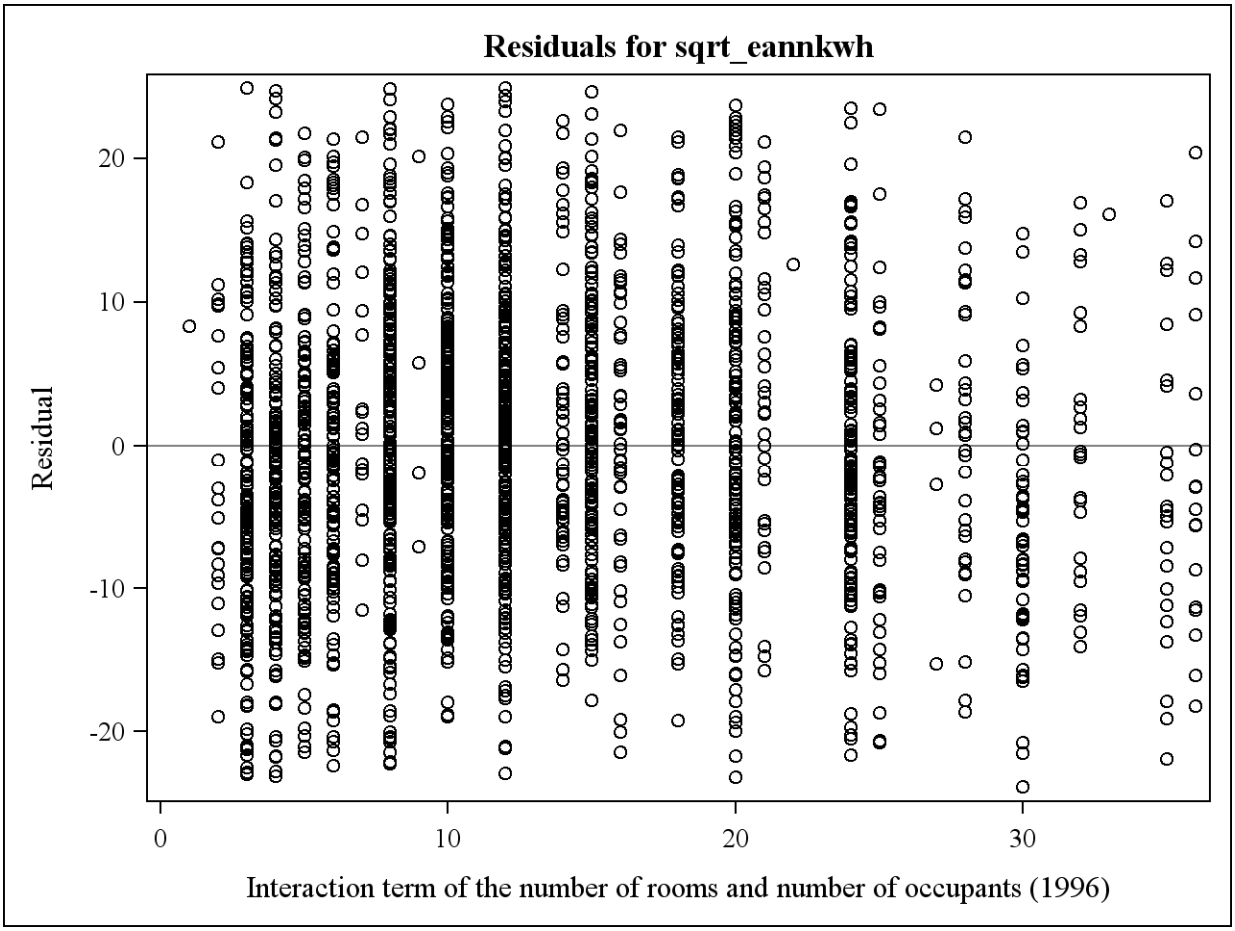
Number of Observations Read	2156
Number of Observations Used	2156

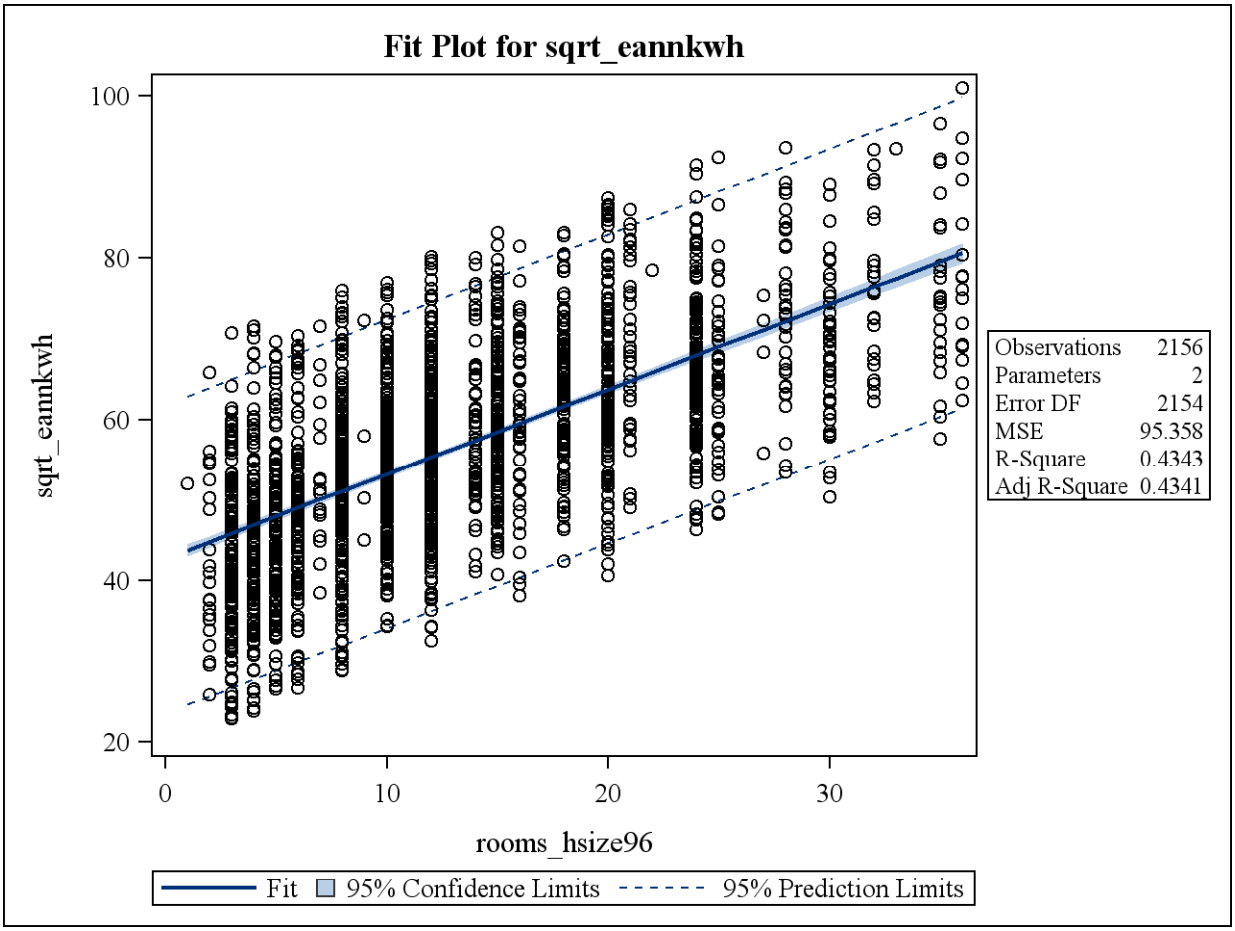
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	157716	157716	1653.93	<.0001
Error	2154	205402	95.35848		
Corrected Total	2155	363119			

Root MSE	9.76517	**R-Square**	0.4343
Dependent Mean	56.33835	**Adj R-Sq**	0.4341
Coeff Var	17.33307		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	42.72071	0.39541	108.04	<.0001
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	1	1.05281	0.02589	40.67	<.0001


Fit Diagnostics for sqrt_eannkwh
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
sqrt_eannkwh
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 2156
Parameters 2
Error DF 2154
MSE 95.358
R-Square 0.4343
Adj R-Square 0.4341


Residuals for sqrt_eannkwh
Residual
20
10
0
-10
-20
0
10
20
30
Interaction term of the number of rooms and number of occupants (1996)


Fit Plot for sqrt_eannkwh
sqrt_eannkwh
100
80
60
40
20
0
10
20
30
rooms_hsize96
Observations 2156
Parameters 2
Error DF 2154
MSE 95.358
R-Square 0.4343
Adj R-Square 0.4341
Fit
95% Confidence Limits
95% Prediction Limits

Running the annual single-level model for the year 2008 (7.2.3)

Number of Observations Read	2121
Number of Observations Used	2121

Descriptive Statistics						
Variable	**Sum**	**Mean**	**Uncorrected SS**	**Variance**	**Standard Deviation**	**Label**
Intercept	2121.00000	1.00000	2121.00000	0	0	Intercept
rooms_hsize96	26807	12.63885	461611	57.92517	7.61086	Interaction term of the number of rooms and number of occupants (1996)
sqrt_eannkwh	120485	56.80581	7242125	187.67399	13.69942	

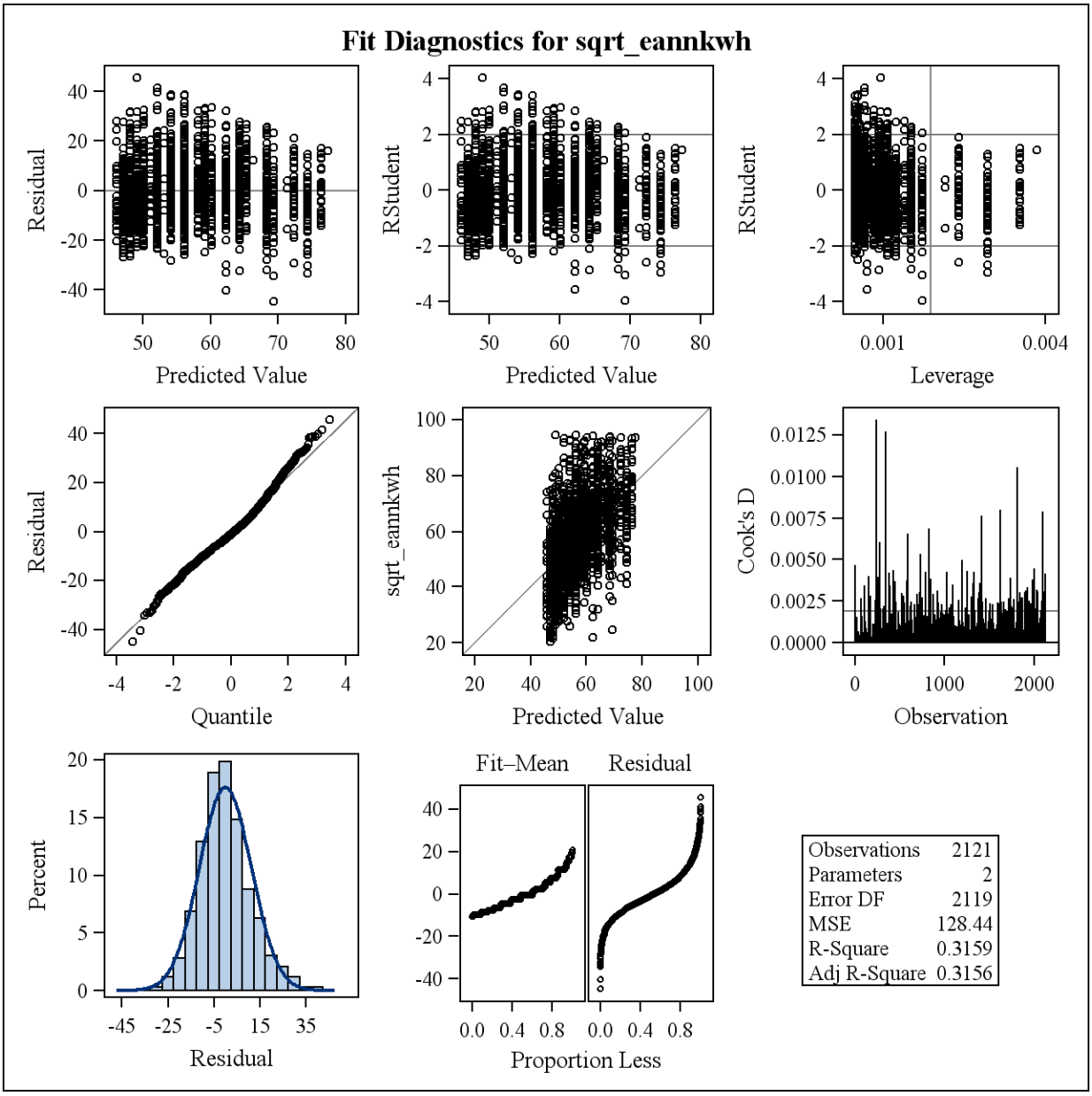
Correlation			
Variable	**Label**	**rooms_hsize96**	**sqrt_eannkwh**
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	1.0000	0.5621
sqrt_eannkwh		0.5621	1.0000

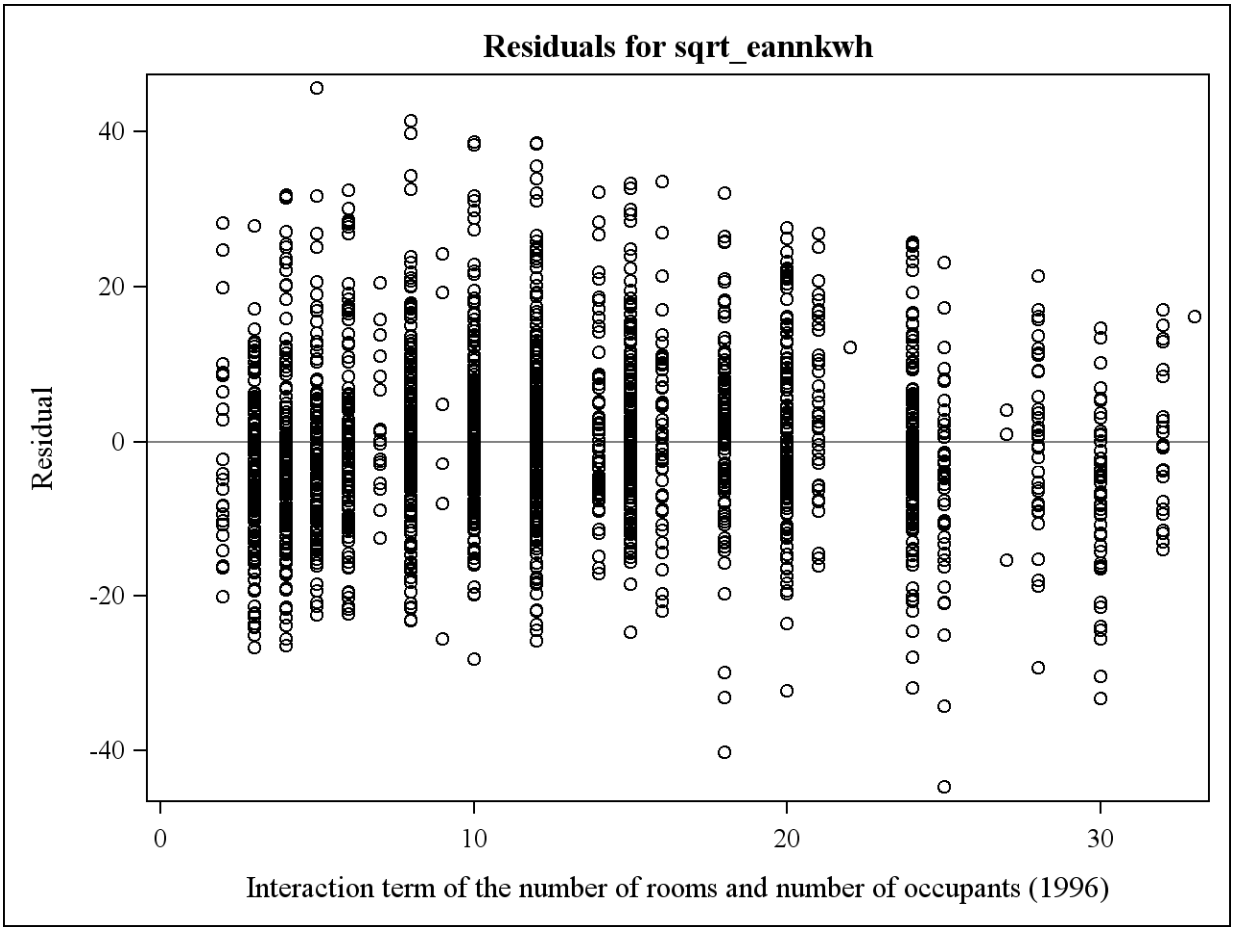
Number of Observations Read	2121
Number of Observations Used	2121

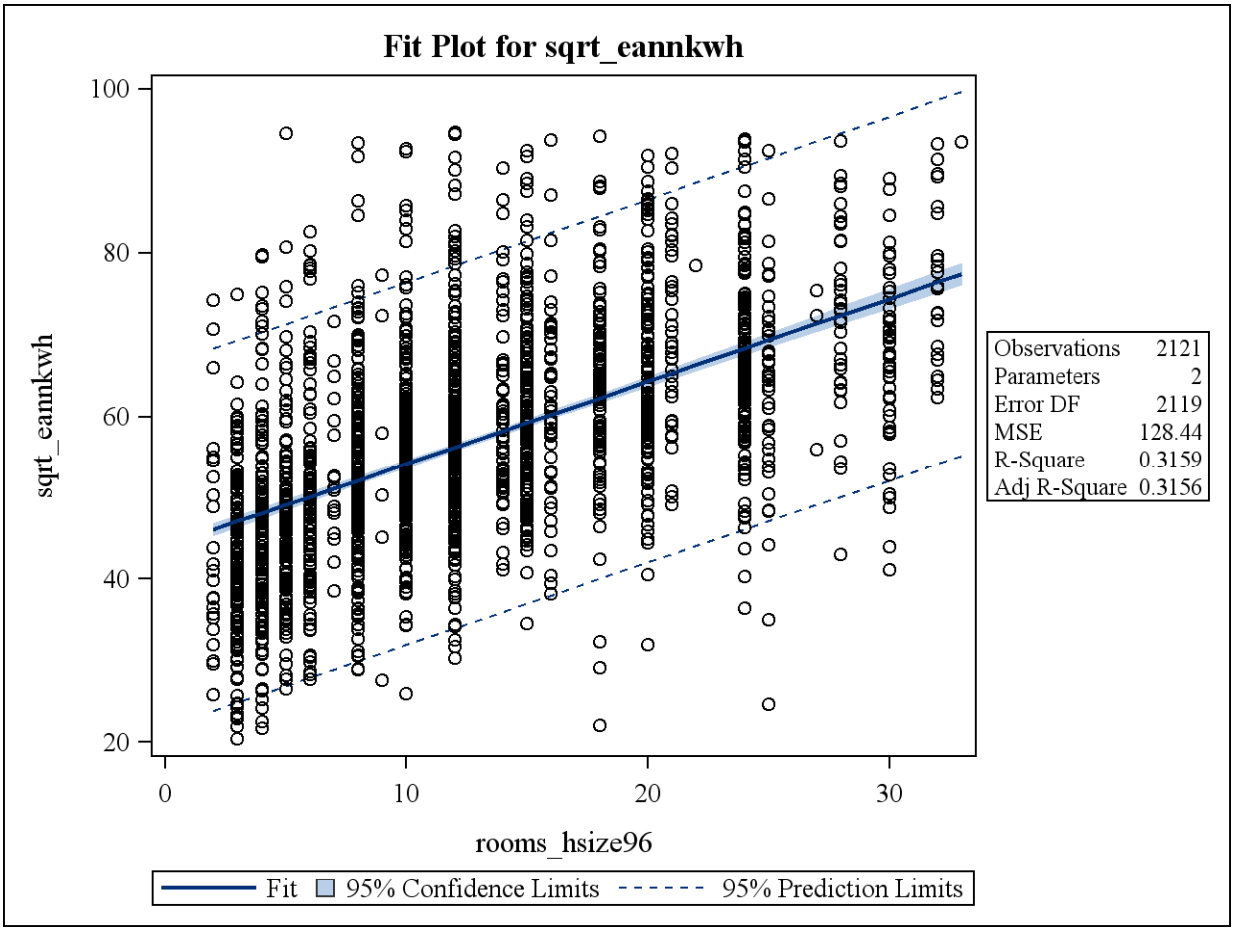
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	125705	125705	978.71	<.0001
Error	2119	272164	128.43987		
Corrected Total	2120	397869			

Root MSE	11.33313	**R-Square**	0.3159
Dependent Mean	56.80581	**Adj R-Sq**	0.3156
Coeff Var	19.95065		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	44.01843	0.47711	92.26	<.0001
rooms_hsize96	Interaction term of the number of rooms and number of occupants (1996)	1	1.01175	0.03234	31.28	<.0001


Fit Diagnostics for sqrt_eannkwh
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
sqrt_eannkwh
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 2121
Parameters 2
Error DF 2119
MSE 128.44
R-Square 0.3159
Adj R-Square 0.3156


Residuals for sqrt_eannkwh
Residual
40
20
0
-20
-40
0
10
20
30
Interaction term of the number of rooms and number of occupants (1996)


Fit Plot for sqrt_eannkwh
sqrt_eannkwh
100
80
60
40
20
0
10
20
30
rooms_hsize96
Observations 2121
Parameters 2
Error DF 2119
MSE 128.44
R-Square 0.3159
Adj R-Square 0.3156
Fit
95% Confidence Limits
95% Prediction Limits

Appendix C: Multilevel modelling parameter estimates

Unconditional means model with supergroups (7.4.1)

Model Information	
Data Set	WORK.WITHMEANS2
Dependent Variable	sqrt_eannkwh
Covariance Structure	Variance Components
Subject Effect	SUPERGROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Dimensions	
Covariance Parameters	2
Columns in X	1
Columns in Z Per Subject	1
Subjects	7
Max Obs Per Subject	361

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
0	1	12857.40337718	
1	2	12826.58753227	0.00000000

Convergence criteria met.

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	SUPERGROUP_CODE	5.2830	3.5179	1.50	0.0666
Residual		168.48	5.9526	28.30	<.0001

Fit Statistics	
-2 Res Log Likelihood	12826.6
AIC (smaller is better)	12830.6
AICC (smaller is better)	12830.6
BIC (smaller is better)	12830.5

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	57.0473	0.9415	6	60.59	<.0001

Unconditional means model for groups and English regions (7.4.2)

Model Information	
Data Set	WORK.WITHMEANS2
Dependent Variable	sqrt_eannkwh
Covariance Structure	Variance Components
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Dimensions	
Covariance Parameters	2
Columns in X	1
Columns in Z Per Subject	1
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	12857.40337718	
1	3	12830.89496685	0.00006497
2	1	12830.52039552	0.00000648
3	1	12830.48614466	0.00000009
4	1	12830.48572009	0.00000000

Convergence criteria met.

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
Intercept	GROUP_CODE	4.5434	2.1074	2.16	0.0155
Residual		168.03	5.9536	28.22	<.0001

Fit Statistics	
-2 Res Log Likelihood	12830.5
AIC (smaller is better)	12834.5
AICC (smaller is better)	12834.5
BIC (smaller is better)	12836.5

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	56.6969	0.6038	19	93.91	<.0001

Model Information	
Data Set	WORK.WITHMEANS_REGION
Dependent Variable	sqrt_eannkwh
Covariance Structure	Variance Components
Subject Effect	gor_code
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Dimensions	
Covariance Parameters	2
Columns in X	1
Columns in Z Per Subject	1
Subjects	9
Max Obs Per Subject	337

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	2156
Number of Observations Not Used	0

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
0	1	17171.85424298	
1	3	17155.28625337	0.00000400
2	1	17155.25750245	0.00000009
3	1	17155.25690046	0.00000000

Convergence criteria met.

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	gor_code	2.2488	1.4422	1.56	0.0595
Residual		166.33	5.0759	32.77	<.0001

Fit Statistics	
-2 Res Log Likelihood	17155.3
AIC (smaller is better)	17159.3
AICC (smaller is better)	17159.3
BIC (smaller is better)	17159.7

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	56.3386	0.5754	8	97.90	<.0001

Including effects of group-level predictors – area classification supergroups (7.4.3)

Model Information	
Data Set	WORK.WITHMEANS2
Dependent Variable	sqrt_eannkwh
Covariance Structure	Variance Components
Subject Effect	SUPERGROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	2
Columns in X	2
Columns in Z Per Subject	1
Subjects	7
Max Obs Per Subject	361

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
0	1	12833.99646610	
1	2	12822.60288055	0.00000003
2	1	12822.60270880	0.00000000

Convergence criteria met.

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	SUPERGROUP_CODE	2.8985	2.3449	1.24	0.1082
Residual		168.48	5.9528	28.30	<.0001

Fit Statistics	
-2 Res Log Likelihood	12822.6
AIC (smaller is better)	12826.6
AICC (smaller is better)	12826.6
BIC (smaller is better)	12826.5

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	56.9555	0.7362	5	77.37	<.0001
super_mean	0.9072	0.4142	5	2.19	0.0801

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
super_mean	1	5	4.80	0.0801

Including group-level predictors of area classification groups (7.4.4)

Model Information	
Data Set	WORK.WITHMEANS2
Dependent Variable	sqrt_eannkwh
Covariance Structure	Variance Components
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	2
Columns in X	2
Columns in Z Per Subject	1
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
0	1	12846.07783097	
1	3	12828.44800577	0.00001896
2	1	12828.34392138	0.00000071
3	1	12828.34031446	0.00000000

Convergence criteria met.

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	GROUP_CODE	3.7514	1.9318	1.94	0.0261
Residual		168.06	5.9558	28.22	<.0001

Fit Statistics	
-2 Res Log Likelihood	12828.3
AIC (smaller is better)	12832.3
AICC (smaller is better)	12832.3
BIC (smaller is better)	12834.3

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	56.7938	0.5703	18	99.58	<.0001
group_mean	0.4821	0.2759	18	1.75	0.0977

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
group_mean	1	18	3.05	0.0977

Including effects of individual household size at the group level (7.4.5)

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Unstructured
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	4
Columns in X	2
Columns in Z Per Subject	2
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Estimated G Correlation Matrix				
Row	**Effect**	**Group_ Code**	**Col1**	**Col2**
1	Intercept	1.1	1.0000	-1.0000
2	hsize	1.1	-1.0000	1.0000

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	GROUP_CODE	6.3189	2.6230	2.41	0.0080
UN(2,1)	GROUP_CODE	-0.1055	0.09154	-1.15	0.2491
UN(2,2)	GROUP_CODE	0.000754	0.006424	0.12	0.4533
Residual		94.3662	3.3769	27.94	<.0001

Fit Statistics	
-2 Res Log Likelihood	11917.8
AIC (smaller is better)	11925.8
AICC (smaller is better)	11925.8
BIC (smaller is better)	11929.8

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
3	64.09	<.0001

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	55.7584	0.6364	19	87.62	<.0001
hsize	1.0457	0.03027	1588	34.55	<.0001

Including effects of individual household size at the supergroup level (7.4.6)

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Unstructured
Subject Effect	SUPERGROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	4
Columns in X	2
Columns in Z Per Subject	2
Subjects	7
Max Obs Per Subject	361

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	11986.14581040	
1	2	11908.98825724	1637987.8730
2	1	11906.68392682	5872540.2539
3	1	11904.86206642	21770533.422
4	1	11904.56298769	26950905.921
5	1	11904.39030317	29590345.778
6	1	11904.36055433	29922267.020
7	1	11904.34573779	30064731.860
8	1	11904.33834570	30129335.444
9	1	11904.33465392	30159878.032
10	1	11904.33280913	30174696.276
11	1	11904.33188700	30181990.443
12	1	11904.33142603	30185608.616
13	1	11904.33119557	30187410.525
14	1	11904.33107948	30188232.745
15	1	11904.33107327	30189441.316

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
16	1	11904.33107205	30183182.146
17	3	11904.33106766	30170086.874
18	1	11904.33105264	30987441.616
19	1	11904.33104301	33059555.503
20	1	11904.33102704	27144771.564
21	1	11904.33031089	3857306293.9
22	5	11904.32991119	1324616274.7
23	33	11904.32944460	4658085027.7
24	18	11904.32981983	330441605.97
25	17	11904.32942601	1873385149.9
26	1	11904.28582085	2.6408048E16
27	22	11904.28555214	9.910962E14
28	26	11904.28389458	3.2794843E15
29	3	11904.26142133	5.0918188E15

WARNING: Stopped because of too many likelihood evaluations.

Covariance Parameter Values At Last Iteration		
Cov Parm	**Subject**	**Estimate**
UN(1,1)	SUPERGROUP_CODE	5.7304
UN(2,1)	SUPERGROUP_CODE	-0.1701
UN(2,2)	SUPERGROUP_CODE	0.000689
Residual		95.0307

Including both individual-level and group-level predictors in the multilevel model (7.4.7)

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Unstructured
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	4
Columns in X	4
Columns in Z Per Subject	2
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	GROUP_CODE	1.5556	0.9081	1.71	0.0433
UN(2,1)	GROUP_CODE	-0.01459	0.05381	-0.27	0.7863
UN(2,2)	GROUP_CODE	0	.	.	.
Residual		94.2105	3.3416	28.19	<.0001

Fit Statistics	
-2 Res Log Likelihood	11904.7
AIC (smaller is better)	11910.7
AICC (smaller is better)	11910.8
BIC (smaller is better)	11913.7

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
2	10.90	0.0043

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	55.8649	0.3918	18	142.59	<.0001
group_mean	1.1030	0.1916	18	5.76	<.0001
hsize	1.0520	0.03005	1587	35.01	<.0001
group_mean*hsize	-0.02121	0.01499	1587	-1.41	0.1574

Adding in individual variables that could show differently different models (7.4.7)

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Unstructured
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile

Model Information	
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	4
Columns in X	6
Columns in Z Per Subject	2
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	GROUP_CODE	1.5815	0.9268	1.71	0.0440
UN(2,1)	GROUP_CODE	-0.01126	0.05589	-0.20	0.8404
UN(2,2)	GROUP_CODE	1.75E-19	.	.	.
Residual		94.2748	3.3459	28.18	<.0001

Fit Statistics	
-2 Res Log Likelihood	11907.2
AIC (smaller is better)	11913.2
AICC (smaller is better)	11913.2
BIC (smaller is better)	11916.2

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
2	10.73	0.0047

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	55.7844	0.5211	18	107.04	<.0001
group_mean	1.0945	0.1930	18	5.67	<.0001
suburb	0.1135	0.5358	1585	0.21	0.8323
hsize	1.0201	0.05087	1585	20.05	<.0001
group_mean*hsize	-0.02263	0.01510	1585	-1.50	0.1341
suburb*hsize	0.04842	0.06232	1585	0.78	0.4373

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Unstructured
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	4
Columns in X	6
Columns in Z Per Subject	2
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	GROUP_CODE	1.3897	0.8396	1.66	0.0489
UN(2,1)	GROUP_CODE	-0.01755	0.05135	-0.34	0.7325
UN(2,2)	GROUP_CODE	3.94E-20	.	.	.
Residual		94.2017	3.3428	28.18	<.0001

Fit Statistics	
-2 Res Log Likelihood	11902.8
AIC (smaller is better)	11908.8
AICC (smaller is better)	11908.9
BIC (smaller is better)	11911.8

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
2	9.50	0.0086

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	57.1212	0.7980	18	71.58	<.0001
group_mean	1.0355	0.1903	18	5.44	<.0001
urban	-1.4556	0.8124	1585	-1.79	0.0734
hsize	1.0807	0.09219	1585	11.72	<.0001
group_mean*hsize	-0.02220	0.01569	1585	-1.41	0.1573
urban*hsize	-0.03086	0.09945	1585	-0.31	0.7563

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Unstructured
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	4
Columns in X	6
Columns in Z Per Subject	2
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	GROUP_CODE	1.8089	1.0003	1.81	0.0353
UN(2,1)	GROUP_CODE	-0.03279	0.05618	-0.58	0.5594
UN(2,2)	GROUP_CODE	0.000022	0.006870	0.00	0.4987
Residual		93.3907	3.3465	27.91	<.0001

Fit Statistics	
-2 Res Log Likelihood	11891.2
AIC (smaller is better)	11899.2
AICC (smaller is better)	11899.2
BIC (smaller is better)	11903.2

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
3	13.32	0.0040

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	54.5479	0.8433	18	64.68	<.0001
group_mean	0.9444	0.2061	18	4.58	0.0002
house	1.8628	0.8441	1585	2.21	0.0275
hsize	1.1680	0.1071	1585	10.90	<.0001
group_mean*hsize	-0.01071	0.01516	1585	-0.71	0.4800
house*hsize	-0.1702	0.1106	1585	-1.54	0.1240

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Unstructured
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	4
Columns in X	6
Columns in Z Per Subject	2
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	GROUP_CODE	1.6123	0.9229	1.75	0.0403
UN(2,1)	GROUP_CODE	0.002899	0.05640	0.05	0.9590
UN(2,2)	GROUP_CODE	1.12E-19	.	.	.
Residual		91.1696	3.2355	28.18	<.0001

Fit Statistics	
-2 Res Log Likelihood	11853.3
AIC (smaller is better)	11859.3
AICC (smaller is better)	11859.3
BIC (smaller is better)	11862.3

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
2	11.86	0.0027

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	55.7038	0.6205	18	89.77	<.0001
group_mean	1.1430	0.1921	18	5.95	<.0001
elderly	1.4315	0.6230	1585	2.30	0.0217
hsize	1.3869	0.08207	1585	16.90	<.0001
group_mean*hsize	-0.03068	0.01486	1585	-2.07	0.0390
elderly*hsize	-0.4832	0.08939	1585	-5.41	<.0001

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Unstructured
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	4
Columns in X	8
Columns in Z Per Subject	2
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	GROUP_CODE	1.7486	0.9952	1.76	0.0395
UN(2,1)	GROUP_CODE	0.01783	0.05918	0.30	0.7632
UN(2,2)	GROUP_CODE	0	.	.	.
Residual		93.1123	3.3066	28.16	<.0001

Fit Statistics	
-2 Res Log Likelihood	11892.6
AIC (smaller is better)	11898.6
AICC (smaller is better)	11898.6
BIC (smaller is better)	11901.6

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
2	12.00	0.0025

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	55.3972	0.4714	18	117.51	<.0001
group_mean	0.9805	0.2390	18	4.10	0.0007
prewar	1.4027	0.5218	1583	2.69	0.0073
hsize	1.1645	0.04343	1583	26.81	<.0001
group_mean*hsize	-0.04750	0.02267	1583	-2.10	0.0363
prewar*hsize	-0.2340	0.06183	1583	-3.79	0.0002
group_mean*prewar	0.2394	0.2752	1583	0.87	0.3845
group_m*prewar*hsize	0.02474	0.03109	1583	0.80	0.4263

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Variance Components
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	2
Columns in X	8
Columns in Z Per Subject	1
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	GROUP_CODE	1.7501	0.9939	1.76	0.0391
Residual		93.1208	3.3070	28.16	<.0001

Fit Statistics	
-2 Res Log Likelihood	11892.7
AIC (smaller is better)	11896.7
AICC (smaller is better)	11896.7
BIC (smaller is better)	11898.7

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	55.3981	0.4716	18	117.47	<.0001
group_mean	0.9807	0.2391	18	4.10	0.0007
prewar	1.4015	0.5220	1583	2.68	0.0073
hsize	1.1627	0.04342	1583	26.78	<.0001
group_mean*hsize	-0.04600	0.02267	1583	-2.03	0.0426
prewar*hsize	-0.2330	0.06178	1583	-3.77	0.0002

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
group_mean*prewar	0.2361	0.2752	1583	0.86	0.3910
group_m*prewar*hsize	0.02283	0.03105	1583	0.74	0.4624

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Variance Components
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	2
Columns in X	7
Columns in Z Per Subject	1
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
Intercept	GROUP_CODE	1.8607	1.0221	1.82	0.0343
Residual		92.7311	3.2920	28.17	<.0001

Fit Statistics	
-2 Res Log Likelihood	11880.5
AIC (smaller is better)	11884.5
AICC (smaller is better)	11884.5
BIC (smaller is better)	11886.5

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	54.0363	0.7215	18	74.89	<.0001
group_mean	0.9756	0.2079	18	4.69	0.0002
house	1.8836	0.7376	1584	2.55	0.0108
prewar	1.0028	0.5273	1584	1.90	0.0574
hsize	1.1231	0.04591	1584	24.46	<.0001

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
group_mean*hsize	-0.02457	0.01540	1584	-1.60	0.1109
prewar*hsize	-0.2149	0.06205	1584	-3.46	0.0005

Model Information	
Data Set	WORK.WITHMEANS96
Dependent Variable	sqrt_eannkwh
Covariance Structure	Variance Components
Subject Effect	GROUP_CODE
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions	
Covariance Parameters	2
Columns in X	6
Columns in Z Per Subject	1
Subjects	20
Max Obs Per Subject	183

Number of Observations	
Number of Observations Read	2156
Number of Observations Used	1609
Number of Observations Not Used	547

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
Intercept	GROUP_CODE	1.6437	0.9517	1.73	0.0421
Residual		93.1202	3.3048	28.18	<.0001

Fit Statistics	
-2 Res Log Likelihood	11888.2
AIC (smaller is better)	11892.2
AICC (smaller is better)	11892.2
BIC (smaller is better)	11894.2

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	55.4141	0.4646	18	119.26	<.0001
group_mean	1.1157	0.1943	18	5.74	<.0001
prewar	1.3078	0.5138	1585	2.55	0.0110
hsize	1.1621	0.04337	1585	26.80	<.0001
group_mean*hsize	-0.03105	0.01522	1585	-2.04	0.0415
prewar*hsize	-0.2403	0.06134	1585	-3.92	<.0001

Appendix D: Office for National Statistics Area Classifications

Extracts from Bond and Insalco (2007)